Journal of Phonetics 52 (2015) 46–57

Contents lists available at ScienceDirect

Journal of Phonetics

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

Research Article

Listeners use intonational phrase boundaries to project turn ends in spoken interaction

Sara Bögels [1], Francisco Torreira [*],[1]

Language and Cognition Department, Max Planck Institute for Psycholinguistics, Nijmegen, The Netherlands

ARTICLE INFO

Article history:
Received 15 May 2014
Received in revised form
8 April 2015
Accepted 23 April 2015

Keywords:
Turn-taking
Turn projection
Intonation
Phrasing
Prosody
Dialogue
Conversation

ABSTRACT

In conversation, turn transitions between speakers often occur smoothly, usually within a time window of a few hundred milliseconds. It has been argued, on the basis of a button-press experiment [De Ruiter, J. P., Mitterer, H., & Enfield, N. J. (2006). Projecting the end of a speaker's turn: A cognitive cornerstone of conversation. *Language*, *82*(3):515–535], that participants in conversation rely mainly on lexico-syntactic information when timing and producing their turns, and that they do not need to make use of intonational cues to achieve smooth transitions and avoid overlaps. In contrast to this view, but in line with previous observational studies, our results from a dialogue task and a button-press task involving questions and answers indicate that the identification of the end of intonational phrases is necessary for smooth turn-taking. In both tasks, participants never responded to questions (i.e., gave an answer or pressed a button to indicate a turn end) at turn-internal points of syntactic completion in the absence of an intonational phrase boundary. Moreover, in the button-press task, they often pressed the button at the same point of syntactic completion when the final word of an intonational phrase was cross-spliced at that location. Furthermore, truncated stimuli ending in a syntactic completion point but lacking an intonational phrase boundary led to significantly delayed button presses. In light of these results, we argue that earlier claims that intonation is not necessary for correct turn-end projection are misguided, and that research on turn-taking should continue to consider intonation as a source of turn-end cues along with other linguistic and communicative phenomena.

1. Introduction

Everyday conversation is characterized by a regular exchange of turns between interlocutors. Although considerable variation occurs in the timing between the end of one speaker's turn and the beginning of the next (De Ruiter, Mitterer, & Enfield, 2006; Heldner & Edlund, 2010), the most frequent turn transitions occur with only slight gap or overlap, regardless of the language (Sacks, Schegloff, & Jefferson, 1974; Stivers et al., 2009). Since turn-taking appears to be smooth in many instances, the question arises how interlocutors manage to time their responses to previous turns. One possible answer is that listeners wait for the offset of speech (i.e., silence of a certain duration) to start their own production; that is, they follow a reactive strategy. However, since speech planning takes at least a few hundred milliseconds (600 ms for picture naming; Indefrey & Levelt, 2004; Jescheniak, Schriefers, & Hantsch, 2003; 1500 ms for simple sentences, Griffin & Bock, 2000), and since the minimal response time for spoken utterances (uttering a pre-planned sound) is about 200 ms (Fry, 1975), reaction to silence is unlikely to be the explanation for at least a large proportion of turn transitions. It seems therefore that listeners often project the end of a turn before its occurrence on the basis of information present in the ongoing turn. In the paper that initiated the modern study of turn-taking, Sacks et al. (1974, 722) already pointed out that it is unlikely that lexico-syntactic information alone will be sufficient for this prediction: “When it is further realized that any word can be made into a ‘one-word’ unit-type, via intonation, then we can appreciate the partial character of the unit-types' description in syntactic terms”. Later studies seemed to substantiate that both syntactic and intonational information is used for turn-end projection (Ford & Thompson, 1996; Gravano & Hirschberg, 2011; Local & Walker, 2012; Wells & MacFarlane, 1998). Against this position, and on the basis of an online button-press study, De Ruiter et al. (2006) have claimed that lexico-syntactic information alone may be sufficient for turn-end projection, and that intonation is neither necessary nor sufficient for this task. The present study explicitly

[*] Correspondence to: Language and Cognition Department, Max Planck Institute for Psycholinguistics, Wundtlaan 1, 6525XD, Nijmegen, The Netherlands.
E-mail addresses: Sara.Bogels@mpi.nl (S. Bögels), Francisco.Torreira@mpi.nl (F. Torreira).
[1] Equal first authors.

0095-4470/$ - see front matter © 2015 Elsevier Ltd. All rights reserved.
http://dx.doi.org/10.1016/j.wocn.2015.04.004

addresses this controversy with a dialogue task and an online button-press experiment in which we control the occurrence of syntactic and intonational phrase boundaries in Dutch polar questions.

Observational studies of English dialogue have identified cues related to turn ends as opposed to turn-medial positions. Duncan (1972) identified several 'turn-yielding cues' that speakers appear to use in American English, including specific intonational inflections, body motion, stereotyped expressions, and syntactic completion, and observed that the chance for a speaker change increases with the number of cues encountered in the speaker's turn. In a conversation-analytic study of over 400 turn transitions in conversations in American English, Ford and Thompson (1996) found that smooth turn transitions only occur at points of simultaneous syntactic, intonational, and pragmatic completeness. More recently, Gravano and Hirschberg (2011) found that both prosodic and lexico-syntactic cues can predict the yielding of the turn after a silence in a game task in American English. Local and Walker (2012) studied all points that were syntactically (and contextually) complete in a telephone call in British English, and focused on non-pitch phonetic cues. Among the identified cues that project the end of the current turn are final lengthening, the absence of segmental reduction, and audible outbreaths. Koiso, Horiuchi, Tutiya, Ichikawa, and Den (1998) found that, in Japanese, both prosodic and syntactic features were related to turn-taking. Dombrowski and Niebuhr (2005) found that phrase-final pitch rises differed in a number of characteristics depending on whether they were produced in a turn-yielding, or a turn-holding context. Turn ends thus appear to be characterized both by completion at the lexico-syntactic and intonational levels. Both kinds of information are therefore potentially available to listeners for projecting turn ends in turn-taking situations. But observational studies cannot provide direct evidence about what information listeners actually use. For instance, listeners might not rely on intonational cues for purposes of turn-taking even when present, and instead focus on lexico-syntactic cues only, or vice versa.

In order to investigate which information listeners use to identify turn ends, several researchers have presented participants with speech fragments, asking them to predict whether the speaker would continue after the fragment or stop speaking. In agreement with the observational studies mentioned above, some of these studies have shown that both syntactic and intonational completion (Caspers, 1998; for Dutch) as well as other prosodic cues (Hjalmarsson, 2011; for Swedish) are used by listeners to identify turn ends in excerpts extracted from recorded conversations. Geluykens and Swerts (1994) presented listeners with utterances from a production experiment, and also found that listeners could distinguish between turn boundaries and non-turn boundaries on the basis of prosodic cues. In contrast, Schaffer (1983), using fragments from conversation, concluded that there is significant variability in listener's use of intonation cues to turn-ends. One should bear in mind, however, that these studies used offline judgments by listeners who were not taking part in the conversation that they judged, and who were not under the timing constraints operating in conversation (i.e., minimization of long gaps, Sacks et al., 1974). For these reasons, these studies leave open the question whether the same types of cues are used online by interactants in a conversation.

De Ruiter et al. (2006) argued that one reason why intonational cues may not be used for turn-end projection is that they occur too late in the current speaker's turn to allow the next speaker sufficient language planning time. To investigate the role of lexico-syntactic and intonational cues in turn-end projection, they conducted an online experiment with isolated turns from a spontaneous Dutch corpus. Participants were asked to listen to each turn and press a button in anticipation of its end. Turns were presented in three different conditions: in their original form, with flattened pitch, and in a low-pass-filtered condition that obscured lexico-syntactic information. Interestingly, it was found that button-press times for stimuli in which the pitch had been flattened were as accurate as button-press times for unmanipulated turns. In contrast, when the words of the turn were made unidentifiable, participants tended to press the button less accurately (i.e., earlier) with respect the end of the turn than when lexical information was preserved. The authors concluded that lexico-syntactic information is necessary for turn-end projection, and that it is possibly sufficient, since intonation did not seem to be necessary for this task in their experiment.

Another recent online experimental study (Keitel, Prinz, Friederici, von Hofsten, & Daum, 2013) used similar stimulus types (dialogs with either normal or flattened pitch) to investigate the role of lexico-syntactic and intonational information in the acquisition of turn-taking skills by German children. Adult and child participants looked at videos of dialogs while their eye movements were tracked, providing potential evidence of anticipation of turn-ends by early switch of gaze to the next speaker. In concordance with De Ruiter et al. (2006), adults did not anticipate turns better in the dialogs with normal pitch than in the dialogs with flattened pitch. Children younger than 3;0 did not reliably anticipate turn-ends, but 3-year-olds could. However, unlike the adults, these children performed better in dialogs including pitch. The authors concluded that children anticipate turns in conversations in an adult-like manner only after they develop a sophisticated understanding of language.

If it is correct that lexico-syntactic information is sufficient for the accurate timing of responses, then the following question remains: how do listeners know, at a given syntactic completion point in the middle of a turn, whether the end of the turn has been reached or not? For example, the sequence of words 'are you a student' could be a complete question, but could also be the beginning of a longer question such as 'are you a student at this university?'. This is an acute problem for lexico-syntactic models of turn-end projection, because, as Sacks et al. (1974) noted, any syntactic phrase can constitute an appropriate full turn, and thus there are many potential turns within actual turns (a problem parallel to the problem of word-recognition, given that most words contain other words; Cutler, 2012: 48–50). One possibility is that the wider discourse context of the turn determines which syntactic completion point is a plausible turn end. Since De Ruiter et al. (2006) used isolated turns extracted from different conversations and presented them randomly to participants, this cannot be the case in their study. In the case of Keitel et al. (2013), on the other hand, this might well have been the case, since participants in this study listened to conversations.

However, it may also be that De Ruiter and colleagues' and Keitel and colleagues' conclusions about the role of intonation in turn-taking are not correct. In order to control for the role of intonation in turn-end projection, both groups of researchers flattened the pitch in their stimuli, but left all other cues to intonational phrasing intact, in particular the lengthening commonly found at the end of intonational phrases (e.g., Wightman, Shattuck-Hufnagel, Ostendorf, & Price, 1992; Gussenhoven & Rietveld, 1992; Turk & Shattuck-Hufnagel, 2007). Since intonational structure, as most other phonological aspects of an utterance, is cued redundantly across different acoustic dimensions, it is possible that intonational phrases were still identified by listeners on the basis of non-pitch cues such as duration and intensity. Moreover, the experimental tasks in De Ruiter et al. (2006) and Keitel et al. (2013), which used overtly manipulated speech in consecutive trials, may have led participants to focus on non-pitch cues in a strategic way that does not reflect their processing mechanisms in everyday conversation. Casillas and Frank (2013) recently tested children between 1 and 7 years of age using a similar task as Keitel et al. (2013). Interestingly, they controlled for the role of prosody by both flattening the pitch and also normalizing syllabic durations in the stimuli, and found that children up to 7-years-old performed worse in anticipating next turns when prosodic cues were left out.

The current study explicitly tests the claim that lexicosyntactic information is sufficient for turn-end projection, and that other phonetic sources of information such as intonation are not necessary for this task. In order to avoid the strategic use of any specific cues by participants, we made an effort to use materials that are not overtly manipulated from the perspective of the listener. In Experiment 1, a research assistant conducted semi-spontaneous (partially scripted) interviews with participants. Her questions fall in two conditions: short (e.g., 'So you are a student?') or long (e.g., 'So you are a student at the university?'). Crucially, both short and long questions feature the same or similar words up to the end of the short question. Using data from these interviews, we specifically aim to address the following issues. First, we investigate whether lexico-syntactic information is sufficient for projecting turn ends by inspecting the timing of the responses to these questions (Experiment 1). Second, we investigate what kinds of prosodic information listeners may have used to project the ends of the questions and time their answers. We do this by comparing the phonetic realizations of the long and short questions produced in Experiment 1. Third, we investigate whether such prosodic information is indeed used by listeners in an online button-press task (Experiment 2) conducted with materials from Experiment 1. We manipulate the intonational properties of these materials via cross-splicing and truncation of a very small proportion of trials per participant (i.e., 2 out of 18), and observe whether and how our manipulations affect the participants' button-presses.

2. Experiment 1

In Experiment 1, we explicitly investigate whether lexico-syntactic information is sufficient for successful turn-end projection. To do this, we present participants with a series of short and long questions, in which the long questions start with the same or roughly the same words as the short ones (e.g., short question: 'So you are a student?' vs. long question: 'So you are a student here at Radboud University?'). If lexico-syntactic information is sufficient for projecting turn ends, as hypothesized in De Ruiter et al. (2006), cases of overlapping speech or false starts around the syntactic completion point in the middle of the long questions should be observed.

2.1. Methods

2.1.1. Participants

Participants were 22 native speakers of Dutch (7 males, 15 females) who were recruited for a separate EEG study. They all gave informed consent and were paid eight Euros per hour.

2.1.2. Materials and design

We constructed eight simple question templates that could be used in an informal dialogue about topics related to our participants' daily lives. Before the experiment, all participants provided some personal information by answering a questionnaire in written form. This information was used to tailor the questions in the experiment to each participant. Table 1 shows example pairs of questions encountered by our participants (i.e., a short and a long version of each). Participants were alternately assigned to two lists. One list contained the short questions with the odd numbers and the long questions with the even numbers in Table 1. This was reversed in the other list.

Table 1
The eight experimental questions. Parts in italics are exemplary and were tailored to each participant by means of a questionnaire they filled out beforehand (see Section 2.1.3).

Item	Short	Long
1	Dus je bent student?	Dus je bent student hier *op de Radboud Universiteit*?
	'So you are a student?'	'So you are a student here *at Radboud University*?'
2	Je studeert dus *psychologie*?	Je studeert dus *psychologie* hier *op de Radboud Universiteit*?
	'So you study *psychology*?'	'So you study *psychology* here *at Radboud University*?'
3	Je bent *eerstejaars*?	Je bent *eerstejaars* student hier *op de Radboud Universiteit*?
	'You are *first year*?'	'You are *first year* student here *at Radboud University*?'
4	Je doet dus aan *basketbal*?	Je doet dus aan *basketbal op donderda*g?
	'So you play *basketball*?'	'So you play *basketball on Thursdays*?'
5	Je hebt wel vaker meegedaan met experimenten?	Je hebt wel vaker meegedaan met experimenten hier op het MPI?
	'You have participated in experiments before?'	'You have participated in experiments before here at the MPI?'
6	Je woont dus *op jezelf*?	Je woont dus *op jezelf* in Nijmegen?
	'So you live *by yourself*?'	'So you live *by yourself* in Nijmegen?'
7	Je hebt dus *een* bijbaan?	Je hebt dus *een* bijbaan *in een supermarkt* naast je studie?
	'So you work on the side?'	'So you work on the side *in a supermarket* in addition to your studies?'
8	Je bent hier met *de fiets*?	Je bent hier met *de fiets* heengekomen *vanochtend*?
	'You came here by *bike*?'	'You came here by *bike this morning*?'

Both long and short versions of the question pairs were syntactically and pragmatically complete. The long version contained the same or roughly the same words as the short version, extended with a syntactic unit including two or more additional words. As can be seen in Table 1, the long version always had a syntactic completion point that corresponded to the end of the short question. From now on, we refer to this point as the *early syntactic completion point*.

2.1.3. Procedure

Participants came to our laboratory for a separate EEG experiment involving quiz questions. Before the EEG setup, all participants filled out the same questionnaire with the information that we needed to construct our target questions (see the Materials and design subsection above). After the EEG setup, participants entered a sound-proof booth, from which they could communicate with the experimenter, a research assistant, via a microphone and speakers. The experimenter told the participant that they would briefly talk about the participant's daily life before the quiz started.

The experimenter, a research assistant blind to the purpose of the study, asked the target questions, tailored to the participant, in the order displayed in Table 1. The questions were available to her on a piece of paper, but she was asked to produce them in a spontaneous way. She did not receive any instructions on how to produce the questions (e.g., regarding intonation, pauses, speech rate). Between the target questions, she asked other (often open) questions that she made up on the spot, which created a spontaneous dialogue and invited the participant to talk freely (see the Appendix for a transcript of part of one of the dialogs). The dialogs lasted about 5–10 minutes and were recorded with one microphone in the recording booth.

2.1.4. Measurements

For each item, we made two measurements by listening to the recorded audio and inspecting waveforms and spectrograms using Praat software (Boersma & Weenink, 2014). First, we measured the duration from the end of the question to the onset of the response. Second, we measured the duration from the early syntactic completion point to the response. Because both questions and answers were recorded with the same microphone, it was not always possible to measure the duration of overlaps between questions and answers as accurately as non-overlaps. In ten cases of slight terminal overlap between the answer and the last segment of the question, the timing was coded as 0 ms. In cases of longer overlap (only three cases), the overlap duration was estimated as accurately as possible. Two items were excluded from the analysis because the participant coughed or swallowed before the answer. Two other items containing a disfluency in the question were also excluded. Our final dataset consisted of 172 question–answer transitions.

2.2. Results and discussion

If, as De Ruiter et al. (2006) predict, all that participants need to accurately predict turn ends is the lexico-syntactic information in the signal, at least some cases of overlap or false starts around the early syntactic completion point in the long questions should be observed. In our data, however, all answers to long questions except three (i.e., 96%) occur either after the actual end of the long questions or just before in overlap with the last syllable. Regarding the three cases of non-terminal overlap, we observed that they occurred much closer to the actual end of the question than to the early syntactic completion point (their overlap durations were 401, 78, and 148 ms, while their offset from the early syntactic completion point amounted to 2121, 1025, and 468 ms, respectively). Thus, all answers to long questions, including the three cases of non-terminal overlap, occurred closer in time to the actual end of the question than to the early point of syntactic completion. This suggests that the timing of the answers to long questions was not planned with reference to the early syntactic completion point, but rather to the actual end of the questions.

We also checked the number of turn-transition times below 200 ms, which must be the result of turn-end anticipation (i.e., cannot be due to a reaction to the silence following the end of speech, since vocal reaction times involve minimally 200 ms, Fry, 1975). Such short turn transitions under 200 ms were present in 18% of short questions (15 out of 83), and in 57% of long questions (51 out of 89, of which 10 were cases of terminal overlap). Thus, at the very least, more than half of the answers to the long questions and about one-fifth of the answers to the short questions were most likely produced in anticipation of the turn end. This is an interesting finding, since, as reported above, we did not find any case of non-terminal overlap or false early starts in the answers to long questions anticipating the early syntactic completion point in a similar way.

In summary, the same syntactic completion point was treated very differently by our participants when it coincided with the actual end of a question (i.e., in the short questions), and when it appeared in the middle of a question (i.e., in the long questions), even though both short and long questions consisted of similar words up to that point. Because of this observation, we conclude that lexico-syntactic information alone is not sufficient for listeners to project turn ends accurately, and that other forms of information in the signal are also necessary to correctly identify turn ends.

From these results we cannot tell what other information listeners used to correctly project the end of the questions that they encountered. Plausible candidates include final cues to intonational phrasing (e.g., final intonation contours, final lengthening), and earlier prosodic cues to utterance length (e.g., initial *f*0 scaling, initial speech rate). One reason for considering early and final prosodic cues separately is that the latter may occur too late in the speaker's turn to be of much use as turn-taking cues for the listener (e.g., De Ruiter et al., 2006; Levinson, 2013), since listeners will often need several hundreds of milliseconds to plan the initial words of their turn (Indefrey & Levelt, 2004; Indefrey, 2011). Under this hypothesis, listeners would need to rely on information located earlier in the speaker's turn. We address this issue in the following sections.

3. Acoustic inspection

In this section, we assess to what extent prosodic cues to turn completion were present in the questions of Experiment 1. More particularly, we investigate whether such prosodic cues mostly occurred at the ends of turns, which due to reaction latencies might make them less useful for early turn-end anticipation, or whether they rather occurred earlier in the turns.

As a first step, we inspected the questions in Experiment 1 auditorily and instrumentally using Praat software. We noticed that all questions were produced in one intonation phrase, and that most ended with rising intonation (154 out of the 172 questions). The vast majority of questions exhibited one or more initial high pitch accent(s) of different saliency, a final low or high pitch accent in the word carrying the sentence stress, and a rise from there to a high boundary tone in the last syllable of the question. All rising questions therefore exhibited either final low-rising or high-rising intonation (L*H-H% and H*H-H% in ToBI-style notation), two intonational patterns previously attested in Dutch polar questions (Gussenhoven & Rietveld, 2000). Fig. 1 illustrates pitch contours aligned with spectrograms for a short and a corresponding long question extracted from our materials. The short question in the example has final low-rising intonation (i.e., with both a low target and a rise to a high target in the final syllable *dent*), while the long question exhibits high pitch accents on the words *student* and *Radboud*, and a final high-rise throughout the syllable *teit*.

After inspecting the questions impressionistically, we systematically compared the short and long questions up to the early syntactic completion point, that is, those parts of the paired short and long questions that had exactly the same words. As explained above, some of the question types contained different words for different participants, since they had to be adjusted to the participant's personal details (e.g., ‘So you play basketball?’ vs. ‘So you play football?’). For this reason, the comparison was performed on only four pairs of question types (1, 5, 6, and 8; see Table 1) for which the words up to the early syntactic completion point were exactly the same between the short and long versions. Moreover, 14 questions produced with either final falling or final rising–falling–rising intonation were excluded from the statistical comparison. The numbers of questions exhibiting these intonation contours were too low to allow for separate categories in statistical comparisons. In total, 74 questions with a homogeneous lexical and intonational make-up were selected for analysis.

We then took several *f*0, duration, and intensity measures in the interval preceding the early syntactic, and sometimes prosodic, completion point (in the words *Dus je bent student* for both short and long questions in Fig. 1 above), and fitted regression models with each of these measures as a response, question length (short vs. long) as the main predictor, and question type (questions 1, 5, 6, or 8) as a covariate. In other words, we compared statistically several phonetic features of the short questions (e.g., *Dus je bent student?*) with those of the corresponding sequences of words when embedded within the long questions (e.g., the words *Dus je bent student* as produced within the long question *Dus je bent student op de Radboud Universiteit?*). The rationale for using these response variables and the results of each test are presented in Sections 3.1 and 3.2.

3.1. Early prosodic cues

By early cues we mean cues present before the word preceding the early syntactic completion point (e.g., in the words *Dus je bent* within the longer sequence of words *Dus je bent student [*…*]?*). We first inspected the *f*0 scaling in the initial words of the questions. Previous research has shown a positive correlation between the initial *f*0 scaling of an utterance and its length (e.g. Cooper & Sorensen, 1981; Fuchs, Petrone, Krivokapíc, & Hoole, 2013), implying that, from a listener's perspective, the initial *f*0 scaling of an utterance could be useful for estimating utterance length and thus for early turn-end projection (Levinson, 2013). It should be noted, on the other hand, that many studies have failed to find a consistent relationship between initial *f*0 scaling and utterance length (e.g., Liberman & Pierrehumbert, 1984; Van den Berg, Gussenhoven, & Rietveld, 1992; Prieto, Shih, & Nibert, 1996). In order to investigate if *f*0 scaling could have served as an early cue in Experiment 1, we measured the average *f*0 of the initial syllable, which was always unaccented, and the maximum *f*0 in the first content word (i.e., *bent*, *hebt,* and *woont*), which often exhibited a local *f*0 maximum. We then fitted three linear regression models with these *f*0 measures

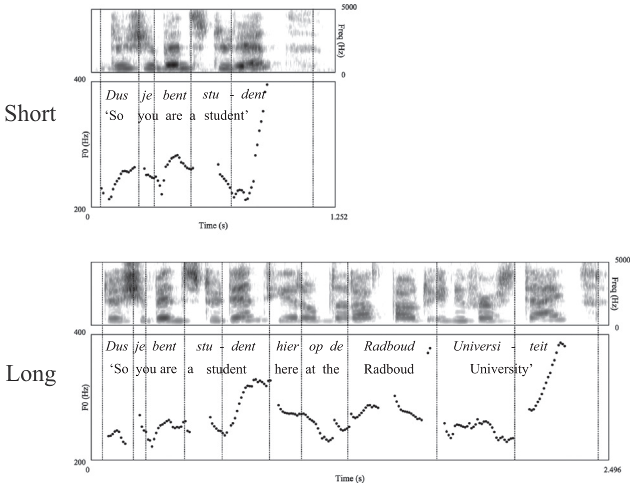

Fig. 1. Examples of intonation contours with aligned spectrograms in a pair of matched short and long questions from Experiment 1 (*Dus je bent student* [*hier op de Radboud Universiteit*]? ‘So you are a student [here at Radboud University]?’). Both examples are drawn to scale in the time dimension. Vertical dashed lines represent the approximate boundaries between specific words/syllables.

(in Hz) and their difference as dependent variables, question length (short vs. long) as the main predictor, and question type as a covariate. No statistically significant differences between the short and long questions were found for any of the three measures, indicating that initial *f*0 scaling was not related to question length in our data.

Given that previous research has found that longer utterances tend to be spoken at a higher speech rate (Nakatani, O'Connor, & Aston, 1981; Quené, 2005; Yuan, Liberman, & Cieri, 2006), we also checked for durational differences between the long and short questions in our data. We examined if the longer questions tended to be spoken at a higher speech rate than the shorter questions. If true, we should see differences in the duration of the interval before the word preceding the early syntactic completion point. A regression model with the duration of this interval as a dependent variable, question length as the main predictor, and question type as a covariate, did not reveal any significant difference between the two groups.

In summary, we did not observe any significant differences in several *f*0 and duration measures taken before the last word preceding the early syntactic completion point. This suggests that the early part of the utterances did not contain salient prosodic cues to their length.

3.2. Final prosodic cues

As said above, we observed that short questions usually ended in low- or high-rising final intonation, both ending in a high boundary tone (H%). In comparison, the early syntactic completion point in the long questions was usually preceded by a word carrying a high pitch accent (H*) only. In both cases, therefore, an *f*0 rise was observed in the last word of the word sequence under inspection. Apart from the fact the *f*0 rise was often noticeably scooped in the case of the low-rising final contours as compared to the rises in the high pitch accents, we observed that the *f*0 rise reached a significantly higher final pitch within the speaker's range when it corresponded to a final boundary tone (H%) than when it corresponded to a phrase-medial high pitch accent (H*). This scaling difference was statistically significant in a regression model with *f*0 final maximum (in Hz) as the dependent variable, question length as the main predictor, and question type as a covariate ($\beta=30.7$; $t=3.88$, $p<.0005$). Importantly, the observed difference in *f*0 scaling was not only statistically significant in terms of group means, but also allowed for a good separation between short and long questions in our data. This is illustrated in the vertical dimension of Fig. 2, which shows final *f*0 height as a function of final syllable duration for short and long questions.

We also observed a difference in the duration of the final syllable in the measured intervals: the last syllables of the short questions were longer than the same syllables when spoken phrase-medially within the longer questions. This was confirmed by a regression model with final syllable duration as the dependent variable, question length as the main predictor, and question type as a covariate ($\beta=.121$, $t=11.16$, $p<.0001$). This finding is consistent with the well-known occurrence of final lengthening at the end of intonational phrases (Turk & Shattuck-Hufnagel, 2007; Wightman et al., 1992, among many others). As for the *f*0 difference above, duration allowed for a very good discrimination between short and long questions, as can be seen in the horizontal dimension of Fig. 2.

Previous studies have shown that turn ends are also characterized by significant intensity drops (Duncan, 1972; Gravano & Hirschberg, 2011). In order to see if this was the case in our materials, we measured the difference between the intensity peaks in the last two syllables of the measured interval. No statistically significant difference was observed in a model with this intensity differential as the dependent variable, utterance length as the main predictor, and question type as a covariate. Although this may have been due to the fact that the microphone used in our recordings was not head-worn, we doubt that our experimental setup would not have captured salient differences in intensity.

In summary, we have identified important differences in *f*0 and duration between our short and long questions in the word located before the early syntactic completion point. These differences are consistent with our initial auditory impressions that this point

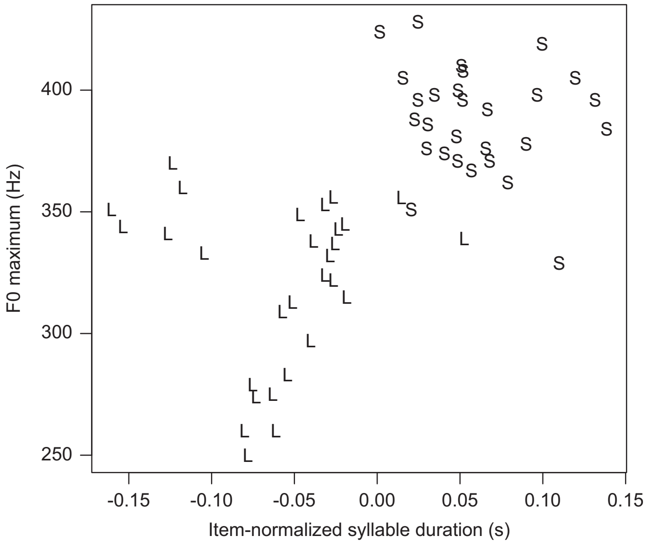

Fig. 2. Item-normalized final syllable duration as a function of *f*0 maximum in the word preceding the early syntactic completion point in question items 1, 5, 6, and 8, for short (S) and long (L) questions.

coincided with the end of an intonational phrase in the short questions, but did not in the long questions. In contrast, we did not find any evidence of early prosodic cues to utterance length in our materials. Nevertheless, we cannot exclude that relevant early cues may have escaped our acoustic inspection, and that these were used by our participants in Experiment 1 for correct turn-end projection. In Experiment 2, we address this issue in a button-press task.

4. Experiment 2

In Experiment 2, we ran the button-press task from De Ruiter et al. (2006) with stimuli created with materials from Experiment 1. We manipulated the presence of early and late prosodic information in stimuli via truncation and cross-splicing of these materials. The main aim of this experiment was to further test De Ruiter and colleagues' claims that lexicosyntactic information is sufficient for turn-end projection by using the same task as in their study. Additionally, we investigated the location of non-lexicosyntactic information in our questions from Experiment 1 that could be used to identify turn ends: early prosodic cues occurring before the last word of the intonational phrase, or late prosodic cues present in the last word of the intonational phrase. Using a simple button press task, we focused on the timing aspect of turn-taking, and controlled variability in response time due to linguistic planning.

4.1. Methods

4.1.1. Participants

Participants were 60 native speakers of Dutch (mean age: 23.1; 32 males and 28 females). They all gave informed consent and agreed to participate in the experiment without any payment (the experiment only lasted about 3 min).

4.1.2. Materials

For the materials, we used three out of the four questions items for which acoustic measurements were available (see Section 3),[2] and for which the early syntactic completion point and the actual end of the utterance displayed a pitch rise (and therefore were minimally acoustically similar). For each of the remaining items, we selected a long and short token representative of their group in terms of *f*0 and duration values. These tokens were used as the original long and short conditions in the experiment. In addition to using these original unmanipulated tokens, we created four manipulated conditions, two short and two long, by means of cross-splicing and truncation (see Fig. 3).

The *fully-replaced* short condition consisted of the initial part of the original long question up to the early syntactic completion point and thus contained both early and late cues from the long question. The *partly-replaced* short condition consisted of the initial part of the original short question up to the word before the early syntactic completion point (potentially containing early prosodic cues), which was cross-spliced with the last word before the early syntactic completion point from the original long question (containing late prosodic cues).

The *fully-replaced* long condition consisted of the complete original short question, which was cross-spliced with the last part of the original long question, from the early syntactic completion point until the end. Thus, this condition contained early and late cues from the short condition. Finally, the *partly-replaced* long condition consisted of the initial part of the original long question up to the word before the early syntactic completion point (potentially containing early prosodic cues), which was cross-spliced with the last word of the original short question (containing late prosodic cues) and the last part of the original long question. Note that overall there were thus three versions of the short question and three versions of the long one, as made clear in Fig. 3.

Random tokens from the other 5 questions from Experiment 1 (short and long versions) were selected from interviews of Experiment 1 and included as question fillers. In addition, 10 statement fillers with final falling intonation (L-L%) were selected from other parts of the interviews or recorded later, and added to the experimental materials to ensure that participants in the experiment would not learn to rely exclusively on final rising intonation as a turn-completion cue.

4.1.3. Design

Participants were assigned to one of six groups (10 participants per group) on the basis of their order of participation. Each group encountered the three possible target items once, each of them in a different experimental condition (one original, one fully-replaced, one partly-replaced). Over all groups, all three target items occurred in all six conditions. We did not present more than two manipulated items to each participant in order to avoid learning effects that would compromise the ecological validity of the experiment. Additionally, each group encountered the same 10 statement fillers as well as 5 additional question fillers from Experiment 1 (forming a total 8 of questions together with the 3 target items, of which 4 were short and 4 long questions). The order of items per participant was pseudo-random, always starting with three statement fillers as practice. To further minimize the chance of learning effects, we controlled the location of the manipulated stimuli within the experiment. The first one always appeared immediately after the practice items, and the other one was the last trial of the experiment.

[2] These were items 1, 6, and 8 from Table 1. The early syntactic completion point in item 5, which corresponded to the end of the word *experimenten* 'experiments' always ended in a pitch fall due to the fact that the H* pitch accent carried by this word was located in its lexically-stressed syllable *men*, not in its last syllable. The end of the utterance, on the other hand, displayed rising final intonation. We decided to discard this item with obvious acoustic differences between the early syntactic completion point and the end of utterance in order to test our hypothesis in a more stringent way.

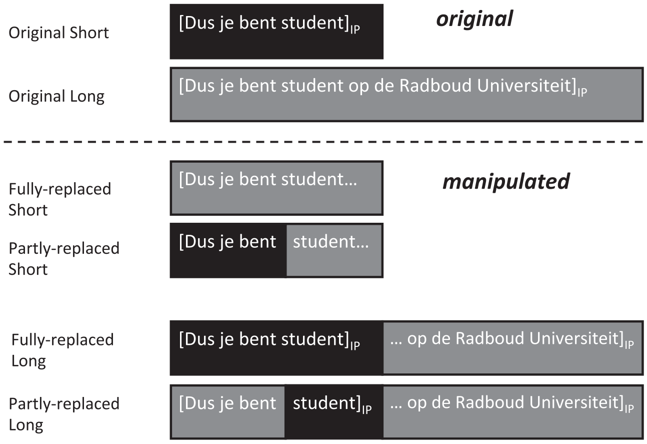

Fig. 3. Different types of target stimuli used in Experiment 2 illustrated with the utterances *Dus je bent student* 'So you are a student' and *Dus je bent student op de Radboud Universiteit* 'So you are a student at Radboud University'. Square brackets represent the beginning ([) and end ($]_{IP}$) in the intonational phrases, whereas the three dots (…) represent points of truncation in the original intonational phrases. Black segments belong to the original short question stimulus, whereas gray segments belong to the original long stimulus.

4.1.4. Procedure

Participants performed the experiment on a laptop and wore sound-cancelling headphones. A trial started with a white fixation cross on a black screen for 1 s, after which the turn started to play over the headphones. Participants were instructed to press a button with their dominant hand when they thought the speaker would be finished speaking. As in De Ruiter et al. (2006), they were asked to try to anticipate that moment instead of reacting to the end of the turn. After pressing the button, the audio file stopped playing immediately and the fixation cross disappeared for 3 s, after which the next trial began.

4.2. Results and discussion

4.2.1. Short questions

We first examine the button presses in the fully-replaced, partly-replaced and original short questions, with the following hypotheses in mind. First, if only lexico-syntactic information is used to project turn ends, as suggested by De Ruiter et al. (2006), all three question types should lead to similar RTs. Second, if final cues to intonational phrasing are used, RTs for manipulated short questions should be longer than those for original questions, since the manipulated short questions do not have an intonational boundary at their end. Third, if participants mainly use early prosodic cues to estimate the length of the questions, only the fully-replaced questions should exhibit longer RTs, since the beginning of the partly-replaced and the original short questions are the same.

Fig. 4 shows average RTs for each short question condition. A regression model with RT as the dependent variable, manipulation type as the main predictor, and random intercepts for participant and item, showed that both kinds of manipulations led to considerably later button presses, with differences of over 300 ms relative to the original baseline (intercept: $\beta=107.1$; partly-replaced: $\beta=332.7$; $t=6.36$, $p<.0001$; fully-replaced: $\beta=312.8$; $t=6.7$, $p<.0001$). Interestingly, the two kinds of manipulations did not produce significantly different RTs. This indicates that it was the prosodic information in the final word of the stimuli that affected the timing of the button presses. Since this final word contains very salient phonetic cues to an intonational phrase boundary in the original stimuli, but does not in the manipulated stimuli (see Section 3), we conclude that intonational phrasing was used by our participants when projecting the end of the short questions.

4.2.2. Long questions

We now compare the button presses of participants in the original long questions containing an early syntactic completion point, and the manipulated long questions with cross-spliced materials taken from the short questions. We address the following hypotheses. First, if only lexico-syntactic information is used for projecting turn ends, participants should exhibit similar patterns of button presses across conditions. Second, if final cues to intonational phrasing are used, button presses should occur close to the early syntactic completion point in the manipulated questions, which exhibit an intonational phrase boundary at that point, but not in the unmanipulated questions. Third, if listeners mainly use early prosodic cues for turn-end projection, they should press the button close to the early syntactic completion point only in the fully-replaced questions.

In order to analyze our data, we defined a one-second window centered at the early syntactic completion point, and counted how many button presses occurred within this window. Fig. 5 shows the results of this analysis. Participants never pressed the button close to the early syntactic completion point in the original long questions, that is, in the absence of an intonational phrase boundary. In contrast, they did press the button in 9 out of 28 cases in the fully-replaced questions, and 8 out of 30 cases in the partly-replaced questions.[3] The estimated probabilities of a button press in both the fully-replaced and partly-replaced questions were statistically different from 0, the estimated

[3] We can only speculate about why participants did not press the button at the syntactic completion point in all manipulated questions. We think that two plausible explanations are (a) that participants heard some more incoming acoustic material before they executed the button press and inhibited their response, and (b) that participants had not planned the button-press at this point yet.

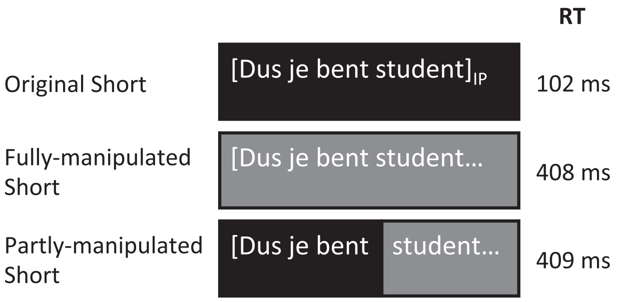

Fig. 4. Average reaction times (RT) per experimental condition for the short question target conditions in Experiment 2.

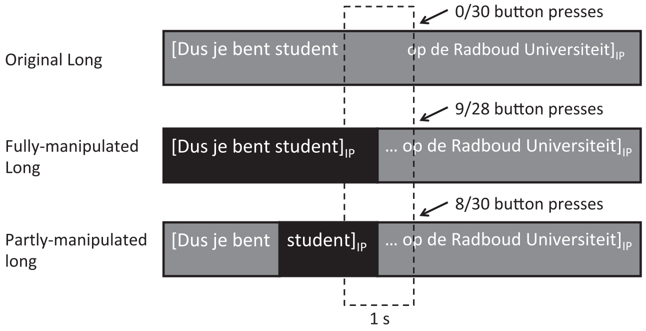

Fig. 5. Number of button presses within 500 ms from the early syntactic completion point for the long question target conditions in Experiment 2.

probability of a button press in the original condition, in separate binomial tests ($p<.0001$ in both cases). Interestingly, as in the case of the short questions, the type of manipulation (partly-replaced vs. fully-replaced) did not seem to affect the results ($p=.52$ in a binomial test comparing these conditions), suggesting that the button presses were affected by the prosodic information in the last word before the early syntactic completion point, and not in earlier words. Just as for the short questions, therefore, it appears that intonational phrase boundaries were used by our participants in order to project the end of turns.

5. General discussion

The previous sections have presented two online experiments aimed at testing two controversial claims about how turn-end projection is achieved in conversation (De Ruiter et al., 2006): first, that lexicosyntactic information is sufficient for the accurate projection of turn-ends; second, that intonation is not necessary for this task. Our results from a dialogue task (Experiment 1) and a button-press task (Experiment 2) run against these claims.

In our experiments, lexicosyntactic information alone was not sufficient for accurate turn-end projection. Participants responded to the same string of words (e.g., “Are you a student”) differently (i.e., with an answer vs. no answer in Exp. 1, and different button-press timings in Exp. 2) based on non-lexicosyntactic information. Our results also indicate that, in the conversational context that we have studied, intonational phrasing was necessary for accurate turn-end projection. In Experiment 1, participants did not respond to syntactic completion points that were not accompanied by an intonational phrase boundary, but they did respond to them when an intonational phrase boundary coincided with the point of syntactic completion. In Experiment 2, we further confirmed that it was the late phonetic cues close to the intonational phrase boundary, rather than other cues potentially present earlier in the signal, that made the detection of turn ends possible. The short original questions ending in an intonational phrase break yielded an average button-press timing of 102 ms relative to the question end, but the manipulated short questions, which did not feature an intonational phrase break at their end, yielded average button-press timings of over 400 ms. Given that average button-press reaction times to auditory stimuli are of the order of 200–400 ms depending on the complexity of the stimulus, we can infer that in the first case participants often reacted to turn-final cues present before the end of the question. In the case of truncated intonational phrases, however, the average latencies of over 400 ms suggest that participants often pressed the button in reaction to the silence following the incomplete utterance.

Our findings therefore indicate that, when timing their response to a turn, participants often attend to late cues to linguistic completion (both lexico-syntactic *and* intonational). Given that planning the production of a single content-word turn requires at least 600 ms (Indefrey & Levelt, 2004), our findings may appear puzzling, since, as claimed in De Ruiter et al. (2006), turn-completion cues that occur near the very end of the interlocutor's turn may come too late for turn-taking purposes (cf. Levinson, 2013). Yet, as we have pointed out, the current findings are in accord with a long tradition of observation of turn-final cues, both prosodic and gestural. The tension between the two views was already clear thirty years ago (Levinson, 1983:301f), but remains unresolved.

The roots of this psycholinguistic puzzle, we believe, lie in over-simplified models of the processes involved in turn-taking. In fact, early projection of content and reaction to turn-final cues are not mutually exclusive (cf. Heldner & Edlund, 2010; Levinson & Torreira, 2015; Torreira, Bögels, & Levinson, 2015). The planning of the content of one's own turn may start midway during the interlocutor's turn once the gist has been grasped, and the very last stage of the production process, actual execution, can be timed locally with reference to late turn-completion cues, provided that the first part of one's turn is ready to be articulated. In this scenario, both syntactic and prosodic completion of the interlocutor's turn can be seen as necessary conditions for the perception of turn completion, since turns only rarely end without them (e.g., in English, any complete utterance, and by extension any turn, must not only be syntactically complete, but also contain at least a nuclear accent and an intonational phrase boundary; cf. Wells & MacFarlane, 1998). It should be noted, however, that these conditions will not be sufficient in many cases, as interlocutors may also want to wait for pragmatic completion (e.g., when a multi-turn unit, such as a story, is in progress). The points at which a turn transition becomes relevant (transition relevance places, or TRPs, in the conversation-analytic literature), can therefore not be determined on the basis of linguistic structure alone, since their projection or detection ultimately depends on pragmatic factors (on this topic, see Selting, 2000 for the differentiation between the end of turn constructional units, or TCUs, and transition relevance places, or TRPs).

Final phonetic cues may also signal that a point of syntactic, prosodic, and pragmatic completion is not an actual turn end (Local & Walker, 2012). For instance, anticipatory coarticulation between a potentially turn-final vowel and the first consonant of an upcoming word (e.g., a bilabial closure at the offset of a phrase-final vowel before a /b/, as in “I do, but […]”) may be interpreted by the listener as a sign that the current speaker will continue. Cues to continuation could also come from tonal coarticulation between phrase-final boundary tones and upcoming tonal targets (e.g., an undershot L% boundary tone before an H* accent in an upcoming word). Another device often used for projecting further talk is that known as the “rush-through” (Schegloff, 1982; Walker, 2010), in which the current speaker locally accelerates her speech rate instead of producing final lengthening at the end of her potentially complete turn, signaling that she is not yielding the floor yet. Languages may also have specific final melodic configurations that act as turn-holding signals, such as the level boundary tone % found in Dutch (Caspers, 2003). From a psycholinguistic perspective, it is likely that listeners often consider phonetic detail (i.e., anticipatory coarticulation) and interactional practices (i.e., rush-throughs, turn-holding melodic configurations) in addition to cues to structural completion (i.e., syntactic and prosodic), and top-down cues to pragmatic completion.

Because our findings are based on question-answer sequences only, one may wonder to what extent they generalize to other conversational contexts. We think it is likely that they do, since prosodic cues to intonational phrasing, such as final lengthening and the presence of characteristic phrase-final tonal patterns, occur in utterances in general, not only in questions. To this respect, it should be noted that Ford and Thompson (1996) observed that smooth turn transitions in their corpus occurred at points of simultaneous syntactic, prosodic, and pragmatic completion in general, not in question-answer sequences alone.

Our proposal that listeners often take into account turn-final cues before articulating their turns is also consistent with the recurrent observation that, in turn transitions, short gaps are more frequent than absences of a gap and short overlaps (Sacks et al., 1974; Jefferson, 1984). In the Corpus of Spoken Dutch, for instance, Heldner and Edlund (2010) found that the most frequent turn transition times lie between 150 and 250 ms. Moreover, of the sixteen observational studies reviewed in Heldner and Edlund (2010), fourteen report positive average turn transition times, and only two report negative values. Given that minimum vocal reaction times observed in the laboratory are of the order of 210 ms for prepared syllables (Fry, 1975), and that turn-final cues tend to occur in the last syllables of turns, a two-stage mechanism allowing for early projection of content and late vocal reaction to turn-final cues predicts the observed distributions better than a model allowing early prediction of both content and timing only. If the timing of turn ends were projected non-locally using early lexico-syntactic prediction, one would expect a most frequent turn transition centered on 0 ms, with variation around this time point due to prediction error. Under that hypothesis, there is no *a priori* reason why a bias towards short gaps should be observed. One may also wonder why speakers would use turn-holding phonetic practices such as the rush-through (Walker, 2010) in a turn-taking system in which listeners focus on lexico-syntactic cues exclusively. Notice that accelerating one's speech rate does not facilitate speech production, but rather makes it more difficult, and that, for this reason, rush-throughs are most likely to be a listener-based phonetic practice. We therefore believe that a production mechanism in which articulation can be launched locally with reference to turn-final completion cues provides a better account not only of our experimental findings, but also of those of past observational studies.

Because turn-end projection is a complex activity, we doubt that any single cue or group of cues will be sufficient or even necessary for the detection of turn ends across the board in all conversational contexts. It is more likely that conversational participants make use of deterministic and probabilistic knowledge about the nature of incoming turns and their situational context in order to project and detect turn ends. In this study, we have shown that intonational phrasing is one of the elements that conversational participants can use in order to project turn ends. Our findings therefore run against recent claims that one does not need to pay attention to intonation in order to understand how smooth turn-taking is achieved in conversation. Accordingly, we believe that future research on turn-taking, both observational and experimental, should consider prosodic phenomena such as intonational phrasing along with other linguistic and communicative phenomena.

Acknowledgments

This work was supported by an ERC Advanced Grant (269484 INTERACT) to Stephen C. Levinson. We would like to thank Annet Veenstra for her assistance during the experiment. We are grateful to Stephen C. Levinson, Marisa Casillas, and Kobin Kendrick for helpful comments on earlier versions of this article. We are also grateful to the members of the IFL project at the Max Planck Institute for Psycholinguistics for extensive discussion of this work. We also thank Carlos Gussenhoven for useful comments about the prosody of Dutch.

Appendix

Transcript of part of one of the interviews from experiment 1. E stands for experimenter, P for participant. Arrows indicate scripted experimental questions asked by the experimenter. Square brackets indicate overlapping speech.

E: →	*Uh je bent eerstejaars?* Eh you are first year?
P:	*Ja* Yes
E:	*En uh bevalt het zo een beetje op die Radboud Universiteit of had je liever* And eh are you enjoying the Radboud University or would you rather have *anders gekozen?* chosen differently?
P:	*Uh nee hoor ik vind de Radboud Universiteit erg leuk.* Eh oh no I like the Radboud University very much.
E: →	*Mooi. OK. Je doet dus aan klimmen in je vrije tijd?* Good. OK. So you do climbing in your spare time?
P:	*Ja.* Yes.
E:	*Wat is dat een bijzondere hobby!* What a special hobby!
P:	*Vind je?* You think?
E:	*Ja, vind ik echt, hartstikke leuk. Is dat uh, doe* Yes, I really think so, very nice. Is that eh, do *je dat op het sportcentrum of doe je dat* you do that at the sports center or do you do that *ook wel eens eh echt in in de bergen of…* also sometimes for real in the mountains or…
P:	*Uh nee ik doe nu een cursus op uh de Radboud* Eh no now I am doing a course at the Radboud *Universiteit gewoon in het sportcentrum* University just at the sports center
E:	*Ja [zo'n uh zo'n ticketuur.]* Yes like a ticket hour.
P:	*[Maar 't is (.) erg leuk om te doen] ja.* But it is (.) very nice to do yes.

References

Boersma, P., & Weenink, D. (2014). *Praat: Doing phonetics by computer* [*Computer program*]. Version 5.3.85. ⟨http://www.praat.org/⟩ Retrieved 19.09.14.

Casillas, M., & Frank, M. C. (2013). "The development of predictive processes in children's discourse understanding". In M. Knauff, M. Pauen, N. Sebanz, & I. Wachsmuth (Eds.), *Proceedings of the 35th annual meeting of the cognitive science society* (pp. 299–304). Cognitive Society.

Caspers, J. (1998). Who's next? The melodic marking of question vs. continuation in Dutch. *Language and Speech*, *41*, 375–398.

Caspers, J. (2003). Local speech melody as a limiting factor in the turn-taking system in Dutch. *Journal of Phonetics*, *31*, 251–276.

Cooper, W. E., & Sorensen, J. M. (1981). *Fundamental frequency in sentence production*. Springer-Verlag.

Cutler, A. (2012). *Native listening*. The MIT Press.

Dombrowski, E., & Niebuhr, O. (2005). Acoustic patterns and communicative functions of phrase-final *F*0 rises in German: Activating and restricting contours. *Phonetica*, *62*, 176–195.

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

Ford, C. E., & Thompson, S. A. (1996). Interactional units in conversation: Syntactic, intonational, and pragmatic resources for the management of turns. In E. Ochs, E. A. Schegloff, & S. A. Thompson (Eds.), *Interaction and Grammar* (pp. 134–184). Cambridge University Press.

Fry, D. B. (1975). Simple reaction-times to speech and non-speech stimuli. *Cortex*, *11*, 355–360.

Fuchs, S., Petrone, C., Krivokapíc, J., & Hoole, P. (2013). Acoustic and respiratory evidence for utterance planning in German. *Journal of Phonetics*, *41*, 29–47.

Geluykens, R., & Swerts, S. (1994). Prosodic cues to discourse boundaries in experimental dialogues. *Speech Communication*, *15*, 69–77.

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

Griffin, Z. M., & Bock, K. (2000). What the eyes say about speaking. *Psychological Science*, *11*, 274–279.

Gussenhoven, C., & Rietveld, T. (1992). Intonation contours, prosodic structure and preboundary lengthening. *Journal of Phonetics*, *20*, 283–303.

Gussenhoven, C., & Rietveld, T. (2000). The behavior of H* and L* under variations in pitch range in Dutch rising contours. *Language and Speech*, *43*, 183–203.

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

Hjalmarsson, A. (2011). The additive effect of turn-taking cues in human and synthetic voice. *Speech Communication*, *53*, 23–35.

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

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

Jefferson, G. (1984). Notes on some orderlinesses of overlap onset. In V. D'Urso, & P. Leonardi (Eds.), *Discourse analysis and natural rhetoric* (pp. 11–38). Padua, Italy: Cleup Editore.

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*, 441–454.

Keitel, A., Prinz, W., Friederici, A. D., von Hofsten, C., & Daum, M. M. (2013). Perception of conversations: The importance of semantics and intonation in children's development. *Journal of Experimental Child Psychology*, *116*, 264–277.

Koiso, H., Horiuchi, Y., Tutiya, S., Ichikawa, A., & Den, Y. (1998). An analysis of turn-taking and backchannels based on prosodic and syntactic features in Japanese Map Task dialogs. *Language and Speech*, *41*, 295–321.

Levinson, S. C. (1983). *Pragmatics*. Cambridge: Cambridge University Press.

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

Levinson, S. C., & Torreira, F. (2015). Timing in turn-taking and its implications for processing models of language. *Front. Psychol.*, *6*, 731 10.3389/fpsyg.2015.00731.

Liberman, Mark, & Pierrehumbert, Janet (1984). Intonational invariance under changes in pitch range and length. In M. Aronoff, & R. Oehrle (Eds.), *Language sound structure*. MIT Press.

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

Nakatani, L. H., O'Connor, J. D., & Aston, C. H. (1981). Prosodic aspects of American English speech rhythm. *Phonetica*, *38*, 84–106.

Quené, H. (2005). Modeling of between-speaker and within-speaker variation in spontaneous speech tempo. In *Proceedings of interspeech 2005* (pp. 2457–2460).

Prieto, P., Shih, C., & Nibert, H. (1996). Pitch downtrend in Spanish. *Journal of Phonetics*, *24*, 445–473.

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

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. (1982). Discourse as an interactional achievement: Some uses of “uh huh” and other things that come between sentences. In D. Tannen (Ed.), *Georgetown University roundtable on languages and linguistics 198; analyzing discourse: Text and talk* (pp. 71–93). Georgetown University Press.

Selting, M. (2000). The construction of units in conversational talk. *Language in Society*, *29*, 477–517.

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

Torreira, F., Bögels, S., & Levinson, S. C. (2015). Breathing for answering: The time course of response planning in conversation. *Frontiers in Psychology*, *6*, http://dx.doi.org/10.3389/fpsyg.2015.00284.

Turk, A., & Shattuck-Hufnagel, S. (2007). Multiple targets of phrase-final lengthening in American English words. *Journal of Phonetics*, *35*, 445–472.

Van den Berg, R., Gussenhoven, C., & Rietveld, T. (1992). Downstep in Dutch: Implications for a model. In G. J. Docherty, & D. R. Ladd (Eds.), *Papers in Laboratory Phonology II: Gesture, segment, prosody* (pp. 335–367). Cambridge University Press.

Walker, G. (2010). The phonetic constitution of a turn-holding practice. In D. Barth-Weingarten, E. Reber, & M. Selting (Eds.), *Prosody in interaction* (pp. 51–72). John Benjamins.

Wells, B., & MacFarlane, S. (1998). Prosody as an interactional resource: Turn-projection and overlap. *Language and Speech*, *41*, 265–294.

Wightman, C., Shattuck-Hufnagel, S., Ostendorf, M., & Price, P. (1992). Segmental durations in the vicinity of prosodic phrase boundaries. *Journal of the Acoustical Society of America*, *91*, 1707–1717.

Yuan, J., Liberman, M., & Cieri, C. (2006). Towards an integrated understanding of speaking rate in conversation. In *Proceedings of interspeech 2006* (pp. 541–544).