PROJECTING THE END OF A SPEAKER'S TURN: A COGNITIVE CORNERSTONE OF CONVERSATION

J. P. DE RUITER
Max Planck Institute for Psycholinguistics

HOLGER MITTERER
Max Planck Institute for Psycholinguistics

N. J. ENFIELD
Max Planck Institute for Psycholinguistics

A key mechanism in the organization of turns at talk in conversation is the ability to anticipate or PROJECT the moment of completion of a current speaker's turn. Some authors suggest that this is achieved via lexicosyntactic cues, while others argue that projection is based on intonational contours. We tested these hypotheses in an on-line experiment, manipulating the presence of symbolic (lexicosyntactic) content and intonational contour of utterances recorded in natural conversations. When hearing the original recordings, subjects can anticipate turn endings with the same degree of accuracy attested in real conversation. With intonational contour entirely removed (leaving intact words and syntax, with a completely flat pitch), there is no change in subjects' accuracy of end-of-turn projection. But in the opposite case (with original intonational contour intact, but with no recognizable words), subjects' performance deteriorates significantly. These results establish that the symbolic (i.e. lexicosyntactic) content of an utterance is necessary (and possibly sufficient) for projecting the moment of its completion, and thus for regulating conversational turn-taking. By contrast, and perhaps surprisingly, intonational contour is neither necessary nor sufficient for end-of-turn projection.*

1. INTRODUCTION. Getting one's timing right is a key problem in speaking. When producing and comprehending speech in conversation, we come under a range of psychological and performance pressures, requiring both speed and temporal accuracy. In the flow of interaction, we run a battery of simultaneous tasks: we are perceiving and processing the speech of others; we are formulating our own utterances in advance; we are simultaneously monitoring the internal timing of our own speech and the timing of our own utterances relative to those of our interlocutors; we are monitoring the content of our own speech and correcting problems if detected; we are monitoring others' responses to our utterances and correcting problems if detected; we are producing and comprehending hand gestures and other bodily actions linked to the speech; and much more besides. Among this rich and urgent flow of perceptual information and motor activity, not only do we work to produce utterances that are well-formed and that achieve the purposes they are designed to achieve (e.g. eliciting information, prompting action, etc.), but we are also working to ensure that the timing and content of our speech production are aligned as seamlessly as possible with those of our interlocutors. Utterances are formulated to fit into sequences of social interaction, and such sequences are characterized by the orderly and finely timed transition of interlocutors between speaker/hearer roles. This is the phenomenon of TURN-TAKING in conversation. For you to produce an irrelevant utterance, or one whose deployment is less than impeccably timed, risks making an 'unusual' contribution, which in turn may cause your interlocutor to infer messages you hadn't intended (e.g. you're getting impatient,

* We wish to thank the Max Planck Gesellschaft for supporting our work. The work of J. P. de Ruiter is also supported by European Union grants IST-2001-32311 and FP6-IST2-003747. In addition, we wish to thank Hanneke Ribberink for her assistance in running the experiments, and Anne Cutler, Nick Evans, Stephen Levinson, Asifa Majid, and Tanya Stivers for their helpful comments on earlier versions of this article. None are to blame for shortcomings of this study—we have not always followed their advice. Finally, we wish to thank Brian Joseph, Norma Mendoza-Denton, and two anonymous referees for their help in making this a better article.

you're becoming hesitant). This constant 'threat to face' (Brown & Levinson 1978) means that there is much at stake in getting the timing of speech production just right.

In their influential and widely cited treatment of the turn-taking problem, Sacks, Schegloff, and Jefferson (1974) point out that to achieve the precise timing that enables us to do no-gap-no-overlap turn transitions, we must be anticipating in advance the precise moment at which a speaker's utterance is going to come to a completion point. This allows us to set the wheels of speech production in motion well before our 'in point' arrives, and in turn (ideally) keep exactly one person talking at all times.

In a typical example of a brief phone exchange shown as a transcript in 1 (Rahman: A:1:VM:(4), supplied by Tanya Stivers), transitions of turns at talk by the two speakers involve virtually no overlap or silence. For every turn, the FLOOR TRANSFER OFFSET (FTO), defined as the difference (in seconds) between the time that turn starts and the moment the previous turn ends, is indicated in square brackets; a positive value indicates a gap (short silence) between the two successive turns, whereas a negative value indicates the amount of overlapping speech.[1]

(1) Mat: 'lo Redcah five o'six one?,
Ver: [+0.15] Hello Mahthew is yer mum the:hr love.
Mat: [+0.13] Uh no she's, gone (up) t'town,h
Ver: [+0.24] Al:right uh will yih tell'er Antie Vera rahn:g then.
Mat: [−0.03] Yeh.
Ver: [+0.13] Okay. She's alright is she.
Mat: [+0.10] Yeh,h
Ver: [+0.07] Okay. Right. Bye bye luv,
Mat: [+0.02] Tara, .h
(End call)

The unyielding imperative of participants in a conversation is to minimize both the amount of speech produced in overlap (i.e. avoid having two or more people speaking at the same time) and the amount of silence between successive turns. Our working assumption is that the one-speaker-at-a-time rule is operative in all informal conversational settings. It is important to understand that to propose such an imperative or 'rule' is not to propose that all talk actually proceeds one speaker at a time. There are constant departures from the rule (overlaps, gaps, and so forth, as is made explicit in the Sacks et al. 1974 model; cf. Schegloff 2000:47–48, n. 1), and these departures can be exploited for functional effect (since they are indeed treated by interactants as departures). These departures may in addition mark differences in personal and cultural style (cf. Tannen 1985). We often encounter informal objections to the Sacks et al. 1974 model, or its ilk, on the basis that in such-and-such a culture or social setting, different standards seem to apply (a common one is that 'in Language/Culture X, people talk in overlap all the time'). Whether such claims are true remains an important empirical question. There are to date no systematic studies of informal conversation that provide counterexamples to the claim of a one-speaker-at-a-time 'design feature' for the regulation of conversational turn-taking (Schegloff 2000:2). Sidnell (2001) conducted an empirical investigation of putative 'contrapuntal' conversation in Antigua (reported by Reisman 1974) but found the data to be compatible with the one-at-a-time model. Thus, as in many domains of linguistic analysis, first intuitions turn out not to be supported by empirical data. We strongly encourage further empirical attempts to describe and ac-

[1] Nonspeech sounds like laughter that occurred at the beginning or end of turns were also assumed to be part of that turn.

count for the distribution of timing of conversational floor transfer in other languages, cultures, and social settings.

The present study begins with the observation that in our sizable data set from Dutch two-party telephone conversations the vast majority of floor transfers have very little gap or overlap. Consider our analysis of the FTO values of all 1,521 speaker transitions in a corpus of phone conversations collected in our laboratory (see below for details on data collection). In Figure 1, the distribution of the FTO values are shown: 45% of all speaker transitions have an FTO of between −250 and 250 milliseconds, and 85% of them are between −750 and 750 milliseconds.

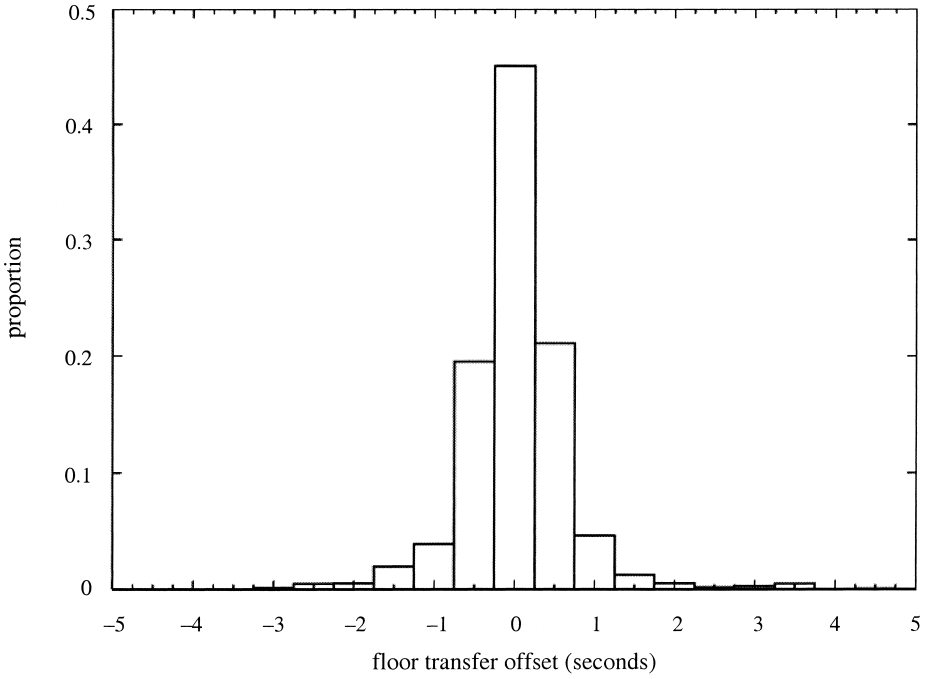

FIGURE 1. Floor transfer offset distribution of 1,521 speaker transitions.

People accomplish this smooth temporal alignment of their talk by anticipating or 'projecting' a moment in time at which transition of speaker/hearer roles will be possible, relevant, and appropriate, allowing them to gear up in advance to begin talking at just the right moment. It is important to emphasize that listeners do not wait until they DETECT the end of a speaker's turn; rather, they ANTICIPATE this moment. An alternative account of turn-taking by Duncan (1974) and Duncan and Fiske (1977) accounts for the regulation of speaker transition by assuming that explicit SIGNALS are exchanged. In the case of turn transitions like the ones in example 1 above, a smooth speaker change would involve a signal from the speaker, followed by a signal from the listener, followed by yet another signal from the speaker (see Figure 1 in Duncan 1974:178). But for the smooth speaker changes under investigation, people simply do not have the time to exchange three signals to negotiate a transfer of floor (see also Cutler & Pearson 1986). Rather, listeners in a conversation must rely on anticipating the moment that the current speaker will complete a turn, making the 'floor' available to be taken up by another speaker. Sacks and colleagues refer to this moment as a TRANSITION

RELEVANCE PLACE (henceforth, TRP), coinciding with the end of a turn constructional unit (i.e. the minimal unit from which turns at talk may be built) produced by a current speaker. Transition relevance in the sense used by Sacks et al. 1974 is a complex notion, where relevance of transition includes reference to the social action being performed by the turn at talk in question (e.g. telling, complaining, requesting, answering). The ability to accurately anticipate the moment of a TRP's occurrence presupposes the ability to project a TURN-COMPLETION POINT (TCP). This ability to project is essential to Sacks and colleagues' 'locally organized' model of turn organization. Only if potential next speakers are able to accurately predict the end of a current speaker's turn can they plan their own turn to be accurately aligned in time with the current speaker's turn to the degree attested empirically. This implies that listeners—as potential next speakers—face the task not only of comprehending the speech they are listening to, but also, in parallel, preparing their next turn and predicting the moment of occurrence of a TCP, where they should be already beginning to speak. In addition, they will need to keep track of issues in the action-sequence structure to assess relevance and/or appropriateness of turn transition, so as to judge when a detected TCP is a TRP. We do not address this here.

Sacks and colleagues do not attempt to solve the question of what is involved psychologically in carrying off accurate end-of-turn projection: 'How projection of unit-types is accomplished, so as to allow such 'no gap' starts by next speakers, is an important question on which linguists can make major contributions. Our characterization in the rules, and in the subsequent discussion, leaves open the matter of how projection is done' (1974:703). We now turn to a discussion of the candidate explanations.[2]

2. PREVIOUS RESEARCH. Several hypotheses have been formulated as to the information sources that listeners use in projecting turn-completion points. Sacks and colleagues (1974) suggested that syntax provides the main projection cue. Other cues that have been suggested in subsequent research include pragmatic information (Ford & Thompson 1996:151), pauses (Maynard 1989), and prosody (Couper-Kuhlen & Selting 1996a and see references below).

Conversation-analytic work on the role of prosody in projection has identified specific intonational contours that often correspond with turn endings. Local, Wells, and Sebba (1985) inspected the turn-final utterance *you know* in a dialect of English called London Jamaican. They showed that the interpretation of *you know* as a turn-yielding cue depended on the prosodic characteristics of the utterance. One of the prosodic cues that corresponded with a turn-yielding *you know* was a falling pitch. Similarly, Local, Kelly, and Wells (1986) found that a rising pitch contour was associated with turn-yielding in Tyneside, another dialect of English, while Wells and Peppé (1996) reported similar findings for Ulster English, a dialect of English that is special because it has a final rising pitch pattern for declaratives.

Especially valuable are those studies that present exhaustive analyses of entire collections of turn transitions in corpora of naturally occurring conversation. Ford and Thomp-

[2] Our consideration of the possibilities is restricted to properties of the speech signal, since we—like Sacks and colleagues (1974)—are working with data from a nonvisual mode of conversation, namely telephone calls. Given our current rudimentary state of understanding of the turn-taking problem, it is important to keep matters simple by minimizing the variables at play. Since conversation typically occurs in a face-to-face situation, it will be necessary in due course to give proper consideration to the role that nonverbal behaviors such as eye gaze and hand gestures play in the organization of turns at talk (Kendon 1967, Goodwin 1981, Bolden 2003, Hayashi 2005, Rossano 2005).

son (1996) analyzed 198 speaker changes in twenty minutes of excerpts from multiparty face-to-face conversation in American English. All excerpts were coded for points of syntactic completion, points of intonational completion, and points of pragmatic completion (i.e. the point where a 'conversational action' is complete, p. 151). Ford and Thompson found that syntactic and intonational completion points frequently coincided, and noted that every pragmatic completion point is an intonational completion point, but not necessarily the other way around. Points in the analyzed conversation where all three types of completion point coincided were termed COMPLEX TRANSITION RELEVANCE PLACES. It is at these complex TRPs, the authors claim, that turn transitions most frequently occur. In their data, nearly three quarters of all speaker changes occur at a complex TRP, and about half of the complex TRPs were accompanied by speaker changes.

Another study, aimed specifically at investigating the role of intonational cues in projection, was done by Caspers (2003), who studied a number of Map Task interactions (Anderson et al. 1991). Like Ford and Thompson, Caspers also found a frequent coincidence of syntactic completion points and intonational completion points. The main result from Caspers's study was that intonational cues are not generally used as turn-yielding signals, but rather that rising pitch followed by a high level boundary tone (H*_% in the 'Transcription of Dutch Intonation' system called ToDI;[3] see Gussenhoven 2005) is used as a turn-KEEPING signal, to 'bridge' syntactic completion points and pauses longer than 100 ms. This contour is characterized by a rise well before the end of an utterance, and a level pitch after this rise until the end of the turn. The idea is that the speaker in such a case is signaling, by using this pitch contour prior to a pause, 'Despite this pause, I'm not finished with my turn'.

In the studies discussed so far, the intonational cues that have been investigated in connection with the projection of turn endings occurred AT the end of turns. This has two implications. First, these cues may be occurring too late in the speech to allow the listener to ANTICIPATE the end of the turn. Even in the case of the H*_% contour described by Caspers (2003), which starts before the end of the turn, the listener has to wait until the end of a turn in order to detect that the pitch contour has a level boundary tone, and not a low boundary tone (H*L%). Second, the observation that certain intonational phenomena COOCCUR with turn endings does not mean that they are used by the listener as anticipatory cues for projection.

The study by Schegloff (1996) is not vulnerable to the first part of this critique, as he identified a pitch cue that occurs before the end of the turn. According to Schegloff, a high pitch peak can signal that completion of the turn will occur at the next syntactic completion point. In this case, listeners would have ample time to anticipate the end of the turn after perceiving the pitch cue. Auer has called this type of explanation the 'filter model' (1996:85). In the filter model, intonation is used by listeners as a filter to decide which syntactic completion points are turn endings, and which are not. However, Auer argues on the basis of data from conversations in German that syntax and intonation are BOTH essential for listeners to determine whether a turn is completed or not.

A few experimental studies have also addressed this issue. Schaffer (1983) used excerpts from conversations and applied low-pass filtering to create versions of these fragments in which the speech was unintelligible, but the intonation was preserved.

[3] To improve readability, we have used the underscore character to denote what is a (meaningful) space in a ToDI string.

She then asked subjects to listen to the fragments and indicate whether they thought the fragments were turn-final or not. Schaffer found that lexicosyntactic information was a much more consistent cue than intonation for end-of-turn detection; subjects listening to unfiltered stimuli showed more (significant) agreement than subjects listening to stimuli from which the lexicosyntactic information, but not the intonational contour, was removed.

Cutler and Pearson (1986) created dialogue fragments by having speakers read written dialogue scripts, and used the recordings of those readings in an experiment in which subjects had to indicate whether the fragments they heard were turn-final or turn-medial. They found some evidence for rising or falling intonational contours signaling turn-finality, but also noted that 'many of the utterances that our listeners found ambiguous also had upstepped or downstepped pitch' (Cutler & Pearson 1986:152). Beattie, Cutler, and Pearson (1982) performed a similar experiment on fragments from interviews with Margaret Thatcher, to find out why she was apparently interrupted so often. They concluded that Mrs. Thatcher's use of sharply dropping intonational contours at positions that were not turn-final often misled listeners into thinking that she was finished with her turn.

To summarize, some studies have argued that syntax is the main information source for accurate end-of-turn projection, while others have suggested that intonational contour also provides essential cues. A complication in evaluating these latter claims is that they rely on an observed COOCCURRENCE of intonational patterns and speaker changes. In natural conversational data, syntactic and intonational structure are correlated (Ford & Thompson 1996, Caspers 2003), which makes it difficult to estimate their relative importance as cues for projection. To reiterate a point already made above, correlation does not by itself imply causation, so establishing a correlation between intonational cues and speaker change does not necessarily mean that the intonational contours are actually USED by listeners. Knowing the location of a TCP in its conversational context does not inform us whether it is intonational contours, syntactic cues, pragmatic cues, pauses, or some combination of these that listeners have used in anticipating the TCP. Other researchers (Beattie et al. 1982, Schaffer 1983, Cutler & Pearson 1986) have tried to approach this problem experimentally by having subjects make judgments on turn-taking issues. But these judgments were off-line, meaning that subjects were not under time pressure to make their judgment. In reality, listeners have to anticipate TCPs in real time, with only fractions of a second to process the speech signal.

3. THE EXPERIMENTAL TASK. We have endeavored to better approximate the conditions under which projection is actually done by having listeners do real-time projection of turn endings, using real conversational data as stimulus materials. We presented fragments of such conversations to subjects and asked them to press a button at the moment they thought that the speaker they were listening to had finished his or her turn. We explicitly instructed our subjects to ANTICIPATE this moment, because we did not want to have the equivalent of a reaction-time experiment where subjects RESPOND to a detected event. This on-line method has the advantage that it mimics the temporal constraints in which projection takes place in nature. Subjects heard the fragments only once and had to anticipate TCPs on the fly.

To be able to tease apart the various candidate cues for projection from among those features that previous descriptive studies have shown tend to cluster at TCPs, our experimental set-up involved manipulation of the speech signal. We needed a way of making a controlled assessment of the relative contribution of three possible cues for

projection: (a) lexicosyntactic information, (b) intonation, or pitch, and (c) amplitude (volume) envelope of the signal. To this end we created stimuli for the on-line projection task by selecting turns at talk from recorded conversations and creating different versions of each, varying the information content in the signal: presence versus absence of lexicosyntactic information; presence versus absence of pitch information; presence versus absence of rhythmical or amplitude-envelope information. The logic is simple. If a certain source of information is used in projection, the absence of that information should have a detrimental effect on the accuracy and/or consistency of the prediction performance of the subjects.

4. STIMULUS COLLECTION. For maximum ecological validity, our stimulus materials were taken directly from recordings of natural conversation. We chose not to construct stimuli using actors, since this may result in omission of natural cues or addition of unnatural ones (see also Cutler & Pearson 1986). However, we had to devise a way not only to get the highest quality audio recordings, but also to avoid cross-talk between the recordings of the two speakers. To accomplish this, we recorded the conversational data using two soundproof cabins. Each cabin had a microphone on a desk in front of the subject and a set of closed headphones. A digital tape recorder was used to record the speech of the subject in cabin A on the left track of a stereo audio recording, and the speech of the subject in cabin B on the right track of the recording. The two speakers could each hear both themselves and their interlocutor in their headphones, giving the effect of a normal phone conversation. The recordings were of top acoustic quality, and there were no traces of the speech of one subject on the recording of the other.

4.1. PROCEDURE. Sixteen native speakers of Dutch participated in the stimulus collection, with eight dyadic interactions (one male-male, three female-female, four female-male). The speakers in each pair were friends. For each dyad, once the two were seated in their respective soundproof cabins, we explained that they could talk to each other via the headphone and microphone, and told them that we were testing our new sound equipment and just needed some random speech samples, so they could freely talk to one another about anything they liked. Each dyad was then recorded for fifteen minutes, resulting in a total of two hours of recorded conversation. The recorded conversations were lively and appeared in all aspects like informal phone conversations between two friends. Thus, there is nothing about the content or composition of these interactions that we would not expect in a natural setting.

5. STIMULUS SELECTION. All recorded conversations were transcribed to a high level of detail, including overlapping speech, turn-beginnings and endings, in-breaths, laughter, pauses, and back-channel signals (Yngve 1970) or 'continuers' (Schegloff 1982) such as *m-hm*.

We then selected a set of turns to use as stimuli in the TCP projection experiments. These turns all contained at least five words. This was to ensure that the subjects in the experiments would have some reasonable amount of signal to base their response on. Short turns contain only limited amounts of lexical and pitch information which could lead to the undesired possibility that we would measure only how fast the subjects were aware that the stimulus had already finished.

We selected 108 target turns and an additional 54 turns for practice purposes. The set of target turns was a random selection from our corpus, although we made sure that the target turns displayed sufficient range with respect to duration and amount of overlap with the turn that followed it in the original conversation (roughly, one third

had some overlap in the original conversation, one third had some silence between the turns (i.e. 'negative overlap'), and one third were almost perfectly aligned). We also made sure that there was a balance between sex of the speaker (50% male and 50% female). The total number of different speakers in the selected turns was fourteen (eight female and six male speakers).

Table 1 presents some descriptive statistics of the target turns. For FTO (see also the discussion of Fig. 1 above), a negative value indicates the duration (in milliseconds) of the overlapping speech with the following turn (which is not audible in the fragment that is presented to subjects in the experiment), whereas a positive value indicates the duration of the silent period between the target turn and the following turn.

	MINIMUM	MAXIMUM	MEAN	STDDEV
Duration (ms)	929	9952	2904	1899
FTO (ms)	−3740	1360	−78	798
Number of words	5	39	13.1	7.85
Mean pitch in Hz	99	300	163	49
(semitones ref 100 Hz)	(0.0)	(19.0)	(8.0)	(4.5)
Std. dev. of pitch within turn (semitones)	0.03	5.77	2.49	1.14
Pitch range within turn (semitones)	0.04	28.80	11.70	5.15

TABLE 1. Descriptive statistics of the selected target turns.

These turns were extracted into individual sound files using the phonetic-analysis program Praat 4.2 (Boersma & Weenink 2004). We extracted only the turn itself, and not the sound from the other channel in which the speech of the interlocutor's subsequent turn had been recorded. Thus, only one speaker would be heard on the resulting stimulus sound file, even if the other speaker was speaking (i.e. in overlap) in the original conversation.

6. PREPROCESSING OF EXPERIMENTAL STIMULI. Five versions were created of every turn fragment. A NATURAL version was the original fragment as is. A NO-PITCH version was created by 'flattening' the pitch (F0) contour by PSOLA resynthesis using Praat 4.2: the pitch was set to the mean pitch value of the original fragment, such that the pitch contour was completely horizontal. A NO-WORDS version was created by low-pass filtering the original fragment at 500 Hz (50 Hz Hanning window). With the NO-WORDS fragments, it is impossible to identify the words, but the original pitch contour remains clearly distinguishable. In a NO-PITCH-NO-WORDS version we obtained fragments with a flat pitch contour AND unidentifiable words by applying both the low-pass filtering and the PSOLA resynthesis. (We applied the low-pass filtering after the PSOLA procedure, because the latter procedure uses information in the higher frequencies for the pitch estimation.) A NO-PITCH-NO-WORDS-NO-RHYTHM version, which we call the NOISE version for brevity, was created by generating a sample of constant noise with the same duration and frequency spectrum as the original fragment. This was achieved by convolving the speech stimulus with white noise. This version was used as a comparative baseline, to see whether the amplitude-envelope information that is still present in the NO-PITCH-NO-WORDS stimuli was an effective cue for projection. All five versions of each fragment were equated in loudness using the sone scale.[4]

[4] The low-pass filtered stimuli generally tend to sound softer because their acoustic energy (which is concentrated in a few critical bands) suffers more from lateral and forward masking than the natural and

Below, we report on two experiments. In the first experiment, subjects were presented with the NATURAL, NO-WORDS, and NO-PITCH stimuli, and in the second, a new group of subjects was presented with the NO-WORDS (this condition was replicated), NO-PITCH-NO-WORDS, and NOISE stimuli. The reason we ran two experiments instead of one experiment containing all five conditions was that it would have complicated the design considerably, leading to a much smaller number of trials in every stimulus/condition design cell. Instead we wanted to collect enough data to derive individual response distributions for every stimulus in every condition.

7. EXPERIMENT 1: NATURAL, NO-PITCH, AND NO-WORDS FRAGMENTS.

7.1. DESIGN. Every subject was presented with three blocks of thirty-six stimuli each, taken from one of the three manipulation groups NATURAL, NO-WORDS, and NO-PITCH. There were six experimental lists; three were permutations of the order in which the blocks were presented, respectively NATURAL (NAT)—NO-PITCH (NP)—NO-WORDS (NW), NW—NAT—NP, and NP—NW—NAT. Within each block, six practice trials were followed by thirty-six randomly selected target stimuli. These were selected such that all 108 target stimuli would be presented in one of the three presented experimental blocks, and none of the 108 stimuli would be presented twice within the same experiment. The remaining three lists were the same as the first three lists, only with the sequential presentation order of the stimuli reversed. These six lists were created in order to counterbalance both potential effects of block order and the order of the stimuli within blocks and over the entire experiment. The order of the practice trials was also reversed for these three lists (but they were, of course, placed before their respective block in the experimental presentation sequence).

7.2. PROCEDURE. Sixty native speakers of Dutch (thirty-nine women and twenty-one men) participated in the experiment. They were assigned randomly to one of the six experimental lists (ten subjects per list). Subjects were seated in a soundproof cabin and given written instruction to listen to the fragments that would be presented to them through closed headphones, and press a button in front of them at the moment they thought the speaker would be finished speaking (Dutch: *is uitgesproken*). The instruction encouraged the subjects to try to ANTICIPATE this moment, and not wait until the fragment stopped playing. They were informed that there were three blocks of trials, and that in some or all of the blocks, the fragments were manipulated acoustically. Subjects were then presented with the six practice trials of the first block. After the practice trials, the first block of stimuli was presented, and after that the same sequence (first practice, then experimental trials) occurred for the second and third block. Every experimental trial consisted of a visual 'countdown' from 3 to 1 presented on the computer screen in front of them, followed by the acoustic presentation of the experimental fragment. An important aspect of the procedure was that as soon as a subject pressed the button, the sound would immediately cut out. This was necessary to avoid giving subjects feedback about their performance. If the subjects still heard the fragment playing after having pressed the button, they might become more conservative, and develop a strategy of waiting for the end of the fragment before pressing the button. Button-press times relative to stimulus onset were recorded by computer. If a subject had not pressed the button by 2,000 ms after stimulus offset, a time-out was recorded.

---

pitch-flattened stimuli. Therefore the sone scale was used instead of the acoustic energy, to compensate for differences in perceived loudness.

8. EXPERIMENT 2: NO-WORDS, NO-PITCH-NO-WORDS, AND NOISE FRAGMENTS. This experiment, with sixty new subjects (forty-four women and sixteen men), differed from experiment 1 only in having the three conditions NO-WORDS (replicated from experiment 1), NO-PITCH-NO-WORDS, and NOISE.

9. RESULTS AND DISCUSSION.

9.1. METHODOLOGICAL VALIDITY. We wanted to know whether the experimental paradigm was tapping into the cognitive process of anticipating TCPs. The following findings demonstrate that indeed it was.

First, for the NATURAL (unmodified) stimuli, response times were very accurate. Subjects were highly skilled at anticipating the moment at which speakers would finish their turn at talk. The average BIAS (defined as response time recorded from stimulus onset minus duration of target stimulus) within the NATURAL condition was −186 ms, indicating that on average, subjects pressed the button 186 ms before the end of the target stimulus. Given that the average duration of the stimuli was 2,904 ms (and the average BIAS only 6% of that), this is remarkably accurate, and reflects what is found in natural conversation. In Figure 2 below, the BIAS of all trials in the NATURAL condition are plotted as a distribution, in the same format as Fig. 1. Similarly to the FTO values in Fig. 1, a negative BIAS value indicates a response before the fragment was finished, and a positive BIAS value a response that occurred after the fragment was finished. A comparison of Fig. 2 with the natural data in Fig. 1 reveals that the projection performance of our experimental subjects had the same mode (at 0) and the same distributional shape as the natural data. Of course, in contrast to the participants in the natural conversations, the subjects in our experiment did not need to prepare and execute a verbal response to the fragment they were listening to, which may explain

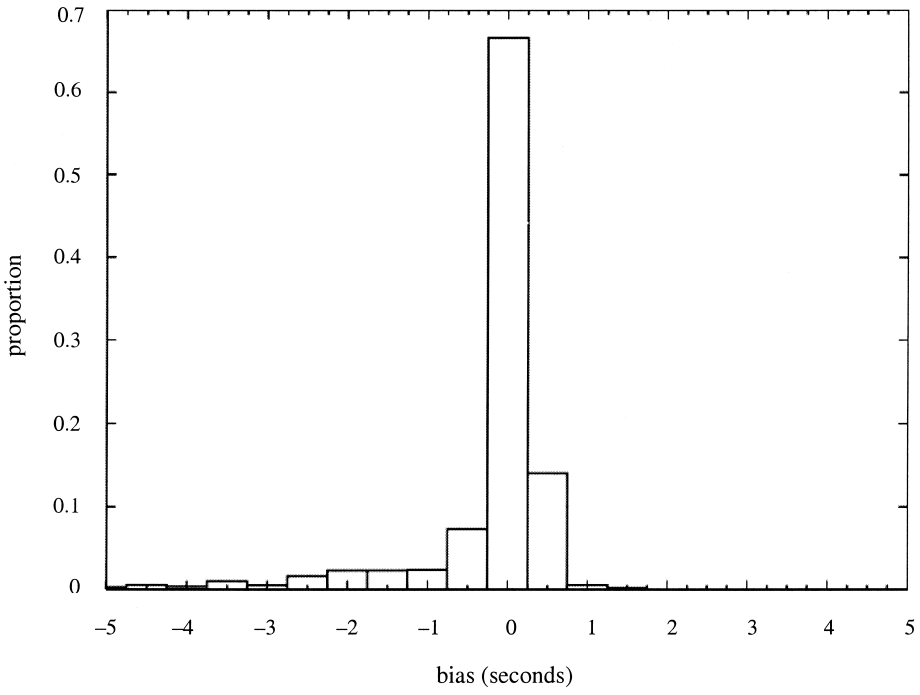

FIGURE 2. BIAS distribution of responses in NATURAL condition.

the higher accuracy (reflected in a higher proportion of values in the FTO = 0 bin) of these button-press projections when compared to the projection performance of the subjects in the real conversations.

Second, the average BIAS was negative in all conditions (BIAS data for the other conditions are discussed below). This indicated that subjects did not, generally, wait for the end of the signal before responding. In some of the trials, the subject responded after the turn was complete, but the negative average BIAS shows that this was not a general strategy. In other words, the subjects were able to ANTICIPATE turn-completion points.

Third, the distribution of responses for many of the longer target turns indicated the existence of multiple TCPs in the stimuli. Given that we had many subject responses to each individual target turn (forty per stimulus in the NO-WORDS condition, and twenty in the other conditions), embedded TCPs that were present in the target stimulus before offset should result in peaks in the response distribution at the earlier TCP as well as the TCP at the end of the stimulus. This is exactly what happened with the majority of the longer stimuli. As a representative example, see Figure 3, in which the waveform of the target turn, the word segmentation, the pitch contour, and the response distributions are displayed on one horizontal time axis. The top panel shows the response distribution for NATURAL (solid line) and NO-PITCH (dotted line) conditions, while the second-to-top panel shows the response distribution for NO-WORDS (solid line) and NO-PITCH-NO-WORDS (dotted line) conditions.

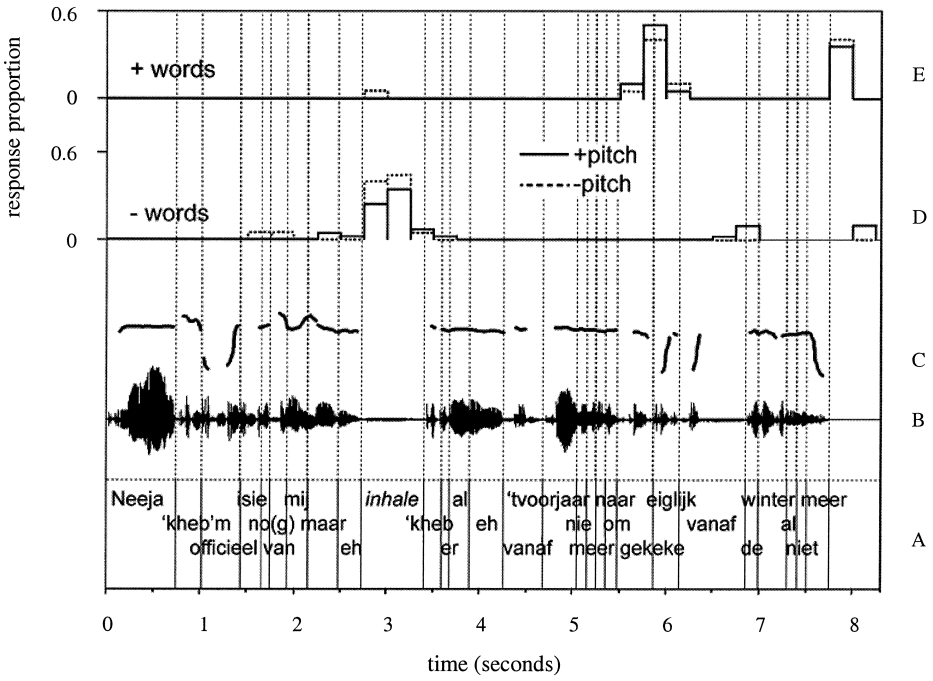

FIGURE 3. Example stimulus with pitch contour and response distributions. From bottom to top, panels represent: A: transcription with each word directly under the piece of speech it occupies, B: sound wave, C: measured F0 (pitch) contour, D: response distribution of the NO-WORDS (solid) and NO-WORDS-NO-PITCH (dotted) conditions, and E: response distribution of NATURAL (solid) and NO-PITCH (dotted) conditions.

The transcription and English gloss of the fragment displayed in Fig. 3 is given in 2.[5]

(2) Neeja 'kheb'm officieel isie no(g) van mij maar eh .hh=
No yes I have it officially is it still mine but eh .hh=
='kheb er al eh vanaf 't voorjaar=
=I have there already eh from the spring=
=nie meer naar omgekeken=
=not anymore cared for=
=eigenlijk vanaf de winter al niet meer.
=actually from the winter on not anymore.

Figure 3 reveals that the response distribution for the NATURAL condition is bimodal, with a peak not only at the end, but also around 2,000 ms before offset, which corresponds with the third syntactic completion point in the turn.[6] (The response distributions of the other two conditions are discussed below.)

9.2. ACCURACY OF SUBJECTS' RESPONSE: BIAS. We now turn our attention to the recorded response BIAS in the different conditions. Figure 4 presents an overview of the average BIAS per condition. In this and all subsequent analyses, ALL recorded valid responses are included, including early responses that may have been caused by embedded TCPs as mentioned above. Due to the nature of the task, longer fragments generally result in a more negative bias than shorter fragments. This is a simple stochastic effect: with the longer fragments the subjects have a longer period during which they could press the button too early than with the shorter fragments. This results in a negative correlation between BIAS and DURATION ($r = -0.636$, $p < 0.001$).[7] Because of this relationship between DURATION and BIAS, all of the statistical tests in this study are performed WITHIN-FRAGMENT. This means that in comparing the different

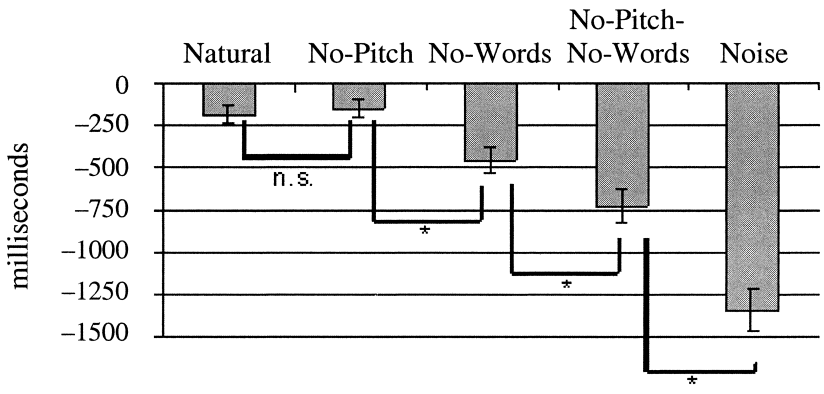

FIGURE 4. Average BIAS of responses per condition.
* indicates statistical significance at the 0.05 level.

[5] A free translation of this fragment: 'No, yes, I have . . . officially it is still mine, but uh actually I haven't given it any attention anymore since spring, in fact since winter.'

[6] We defined syntactic completion points in the same way as Ford and Thompson (1996:143–45).

[7] Interestingly, this relationship is also present in our conversational data. The longer the (first) turn, the more negative the FTO. In our conversational data, the correlation between DURATION and FTO is -0.136 ($p < 0.001$).

conditions, every condition always contains the same 108 stimuli. This way, we excluded the possibility that the relationship between DURATION and BIAS influenced our results.

An ANOVA shows a significant main effect for presentation condition (by subjects: $F1(4,353) = 73.37$, $MSE = 165{,}610$, $p < 0.001$; by items: $F2(4,428) = 64.95$, $MSE = 330{,}938$, $p < 0.001$). A post-hoc analysis (Tukey HSD, $\alpha = 0.05$) indicated that all differences between individual conditions were significant, with the key exception being the difference between the NATURAL and the NO-PITCH condition, which was not significant in either item- or subject-analysis. A general linear model analysis revealed no significant effects of the independent variables SPEAKER-SEX or SUBJECT-SEX on BIAS, nor any significant interactions of SPEAKER-SEX or SUBJECT-SEX with CONDITION. The partial correlation between FTO and BIAS, controlling for DURATION, is not significant in any of the conditions.[8]

The most important finding is that removing pitch information from the stimuli had no influence on projection accuracy. By contrast, removing lexical content by the low-pass filtering procedure in the NO-WORDS condition DID have a strong detrimental effect on projection accuracy.

The significant difference in BIAS between NO-WORDS and NO-PITCH-NO-WORDS suggests that when no lexicosyntactic information is available, pitch can sometimes be used as a 'turn-keeping' cue in the presence of pauses in the signal. Grosjean and Hirt (1996) report a similar finding in experiments where subjects had to predict the end of sentences in a GATING task (Grosjean 1980), noting that 'It is only when higher-level information is no longer present . . . that prosodic information is made available or is called into play' (Grosjean & Hirt 1996:129). In the NO-PITCH-NO-WORDS condition, the main acoustic cue available for projection in this experiment is the amplitude envelope. A silence of a certain duration will, in the absence of any lexicosyntactic information, often be interpreted as the end of the turn. If a turn-keeping pitch contour, for instance the H*_% contour as described by Caspers (2003), is present, this may delay the response. Inspection of the response distributions in the NO-PITCH-NO-WORDS condition confirmed that pauses in the speech signal were indeed the main determinant for the response distributions in the NO-PITCH-NO-WORDS condition. In the data from Fig. 3, the response distributions of the conditions without lexicosyntactic information (NO-WORDS and NO-PITCH-NO-WORDS; see the arrow in the D panel) both peaked exactly at the pause in the speech. In Figure 5, another illustrative stimulus is depicted together with the corresponding response distributions. The transcription and English gloss of the fragment is given in 3.[9]

(3) Ah en waren er nog eh (0.5 s) bijzondere mensen die eraan =
Ah and were there yet eh (0.5 s) special people who there in =
= meededen of nie.
= participated or not.

In the NO-PITCH-NO-WORDS condition, the majority of the responses occur around the 500 ms pause (at about 2.3 s). In contrast, in the NO-WORDS condition, where the pitch contour is still present, the majority of responses occur at the end of the fragment. This indicates that having access to the pitch contour prevented the subjects in the NO-WORDS condition from responding to the pause, whereas the sub-

[8] Partialing out the effect of DURATION is necessary in this analysis, because FTO is negatively correlated with DURATION ($r = -0.207$, $p < 0.05$) for the fragments used in the experiment.

[9] Free translation: 'ah, and were there uh any special people who participated in it or not?'

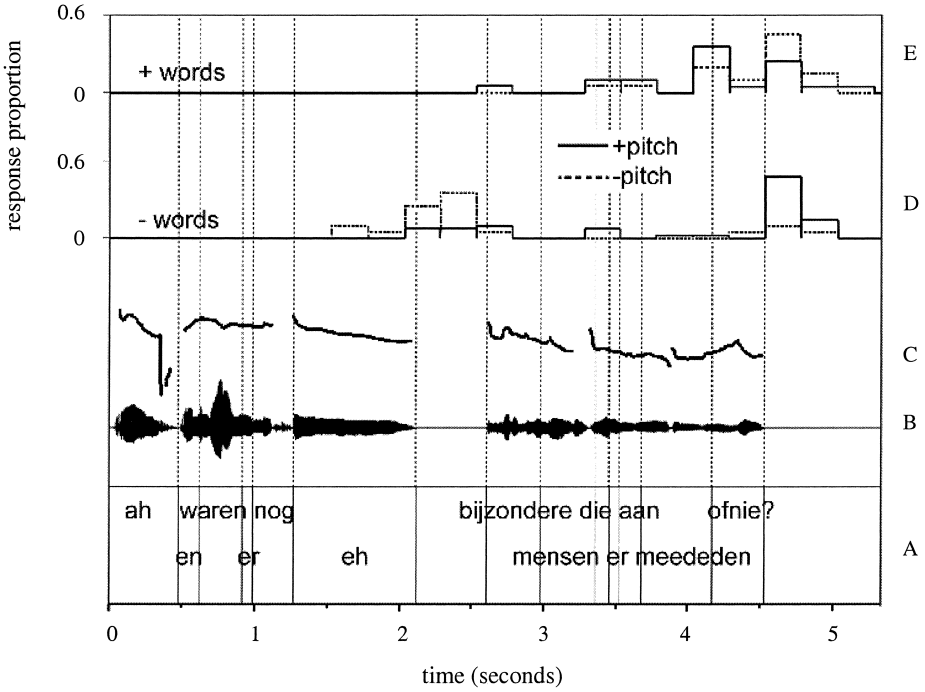

FIGURE 5. Example stimulus with pause. From bottom to top, panels represent: A: transcription with each word directly under the piece of speech it occupies, B: sound wave, C: measured F0 (pitch) contour, D: response distribution of NO-WORDS (solid) and NO-WORDS-NO-PITCH (dotted) conditions, and E: response distribution of NATURAL (solid) and NO-PITCH (dotted) conditions.

jects in the NO-PITCH-NO-WORDS condition were 'fooled' by the pause. The high/level pitch contour serves as a turn-keeping cue to 'override' pauses in the signal (cf. Caspers 2003).

Lexical information can also 'override' pauses in the speech. In the fragments in both Figs. 3 and 5, the pauses are preceded by the word *eh*. Clark and Fox Tree (2002) investigated the function of *uh* (the English equivalent of the Dutch *eh*) in conversation. Although their analysis did not address speaker changes in conversational turn-taking, Clark and Fox Tree raise the possibility that English *uh* could be used as a turn-keeping cue, that is, as signaling 'I'm having trouble producing this utterance but I'm not finished yet, so wait' (see also Jefferson 1973). Our investigation provides strong support for a turn-keeping function of *eh* in conversation. In our stimulus set, there are fifteen fragments containing pauses longer than 200 ms, and ten of them (two thirds) are immediately preceded by *eh*. Furthermore, in the NATURAL and NO-PITCH conditions (where the *eh* could be heard and identified), the likelihood of a response peak at a pause without a preceding *eh* was twice as high as the likelihood of a peak in the response distribution at a pause that WAS preceded by *eh*. In other words, in the conditions where *eh* was identifiable, its presence prevented the pause from being interpreted as a TCP.

The finding that the removal of pitch information does not have any effect on projection accuracy seems, at first sight, to be at odds with the finding by Beattie and colleagues (1982), who investigated why Margaret Thatcher was interrupted so often

during interviews. They found that a rapidly falling pitch in her voice led her interviewers to 'mis-project' and take up the next turn, although apparently Mrs. Thatcher had not yet finished. This shows that under certain circumstances, it IS possible for pitch contours to suggest the presence of a TCP (see also Schegloff 1996). We inspected the individual response distributions for all of our experimental stimuli and located seven fragments for which the mode of the response distribution for the NATURAL data precedes the mode of the NO-PITCH distribution by 250 ms or more. So for these cases, when the stimulus contained pitch information, subjects responded at least 250 ms EARLIER than when the stimulus did NOT contain pitch information. In all of these seven cases, there was either a noticeable pitch rise (two cases, mean pitch velocity +22 semitones/s) or fall (five cases, mean pitch velocity −12 semitones/s), and these pitch contours occurred at syntactic completion points. Both in the Beattie et al. 1982 and our own studies, a rapidly changing pitch could mislead listeners into projecting the end of the turn at an early syntactic completion point. In our study, the presence of these cues led to early responses to these fragments in the NATURAL condition.[10] Intriguingly, these fragments did not result in early responses in the NO-WORDS condition, where the subjects had full access to the pitch contour. If rapidly falling or rising pitch contours BY THEMSELVES indicated the presence of TCPs, they should have led to early responses in the NO-WORDS condition. The fact that they did not means that these pitch cues indicate the presence of a TCP only if lexicosyntactic information is also available (as was the case in all of the conditions in the study by Beattie and colleagues). This suggests that lexicosyntactic information is, again, the crucial source of information, even in the case of 'mis-projections'.

A final BIAS effect to discuss is that even in the NO-PITCH-NO-WORDS condition, projection accuracy was still better than chance, chance performance being represented by the NOISE baseline in which the subjects had absolutely no information for projecting. This indicates that the rhythmical properties of the fragments in the NO-PITCH-NO-WORDS stimuli still contained some cues for projection.

Our main finding from the analysis of BIAS is that removing the pitch contour from the original stimulus did not have an effect on subjects' projection performance. By contrast, removing the lexicosyntactic content (while retaining the pitch contour) did significantly deteriorate projection performance. In those cases where the lexicosyntactic information was removed from the speech, intonation was of some help in signaling that longer pauses in the speech were not turn endings.

9.3. CONSISTENCY OF SUBJECTS' RESPONSES: ENTROPY. Another way to investigate the relative role of pitch and lexical information in projection, in addition to the analysis of the BIAS presented above, is to inspect the CONSISTENCY with which subjects responded to the stimuli in different conditions. That is, we do not ask how ACCURATE they were, but rather how much AGREEMENT there was among subjects with respect to when they responded. The subjects may have been 'wrong' in their projection compared to the original interlocutor, but the higher the agreement among them, the more likely it is that they have used the cues present in that particular condition. A straightforward way to estimate this level of agreement from the data is to compute the standard deviation of all responses for every stimulus/condition pair. This, however, has a serious disadvantage. As has been shown in the analysis above, there can be multiple TCPs

[10] Reanalysis of BIAS excluding the seven stimuli that contained these pitch cues led to an identical pattern in the overall results.

in one turn fragment, leading to multimodal response distributions. The further apart these TCPs are in time, the higher will be the standard deviation of the responses, even though the agreement with respect to the perceived location of those multiple TCPs might have been high. To circumvent this problem, we computed for every stimulus/condition pair the ENTROPY as defined by Shannon (1948),[11] a measure of uncertainty that does not have the disadvantage mentioned above. The entropy measure is insensitive to the actual temporal distance between these multiple modes. However, it is sensitive to the level of agreement. If all responses were to occur within the same interval, the entropy would be zero. The more the responses are distributed over different intervals, the higher the entropy is. Therefore, the entropy gives us a reliable estimate of the amount of agreement among subjects, without being biased by the temporal distance between possible multiple TCPs in the fragments.

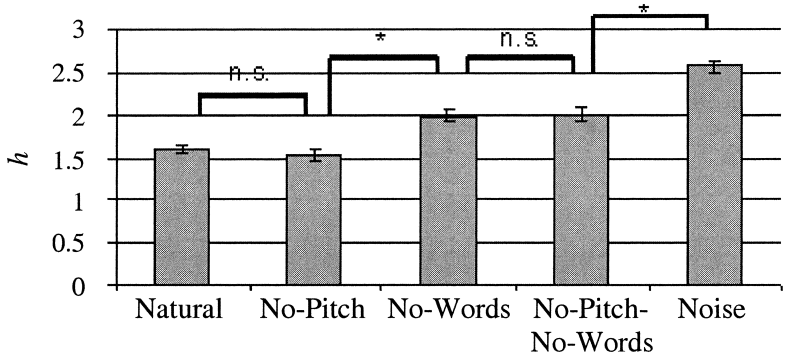

FIGURE 6. Shannon entropy of responses per condition.
* indicates statistical significance at the 0.05 level.

In Figure 6, the average Shannon entropy (using a bin-width of 250 ms) is shown for every condition. Again, there is a main effect of condition ($F2(4,428) = 43.79$, $MSE = 0.206$, $p < 0.001$).[12] A post-hoc analysis (Tukey HSD, $\alpha = 0.05$) indicated that all differences between conditions were significant with two key exceptions: the difference between NATURAL and NO-PITCH and the difference between NO-WORDS and NO-PITCH-NO-WORDS. These differences were not significant.

This result supports the above interpretation of the BIAS data that the removal of pitch information from the turn fragments does not have a detrimental effect on projection performance: the consistency of subjects' responses was unaffected by the removal of pitch information, both for the conditions with and without lexicosyntactic information. A similar conclusion was reached by Schaffer (1983:253), who concluded that 'Syntactic and lexical characteristics appear to be used much more consistently as cues to turn status'. However, Schaffer did not have an experimental condition equivalent to our NO-PITCH condition, where the lexicosyntactic information is present but the intonational contour is not. Therefore, subjects listening to the unfiltered fragments in Schaffer's experiment always had MORE information available than those listening to the filtered

[11] Entropy $h = -\sum_{i=1}^{N} p_i \log(p_i)$, where N is the number of bins, and p_i is the proportion of samples in bin i.

[12] Because entropy distributions can be computed for every stimulus only over entire response distributions, only a by-item statistical analysis could be performed here.

fragments, and this could be an alternative explanation of her finding. Our study ruled out this alternative explanation by varying the presence of intonation and lexicosyntactic information independently.

The entropy results also provide support for the above explanation of the difference in BIAS between the NO-WORDS and the NO-PITCH-NO-WORDS conditions. If subjects did perceive the pitch contours present in the NO-WORDS fragments to be turn-keeping cues when there was a pause in the fragment, this should have affected the BIAS (as it did) but not the agreement about where the TCP was. Subjects in the NO-PITCH-NO-WORDS condition may have been less accurate (hence the larger absolute BIAS), but their level of agreement was nevertheless the same, because they were all 'fooled' in the same way by the same pauses.

10. CONCLUDING DISCUSSION. In our experimental simulation of the conditions of end-of-turn projection in Dutch conversation, subjects demonstrated the same remarkable ability as speakers in natural conversation to accurately anticipate the moment of end-of-turn completion. Different experimental conditions revealed that in order to achieve this accuracy of projection, what they needed was the lexicosyntactic structure of the target utterances. When we removed pitch contour from the target utterance, leaving the lexicosyntax intact, this had NO EFFECT on hearers' ability to accurately project turn endings. And when we removed the lexicosyntactic content, leaving the original pitch shape in place, we saw a dramatic decline in projection performance. The conclusion is clear: lexicosyntactic structure is necessary (and possibly sufficient) for accurate end-of-turn projection, while intonational structure, perhaps surprisingly, is neither necessary nor sufficient.

This conclusion raises a host of further questions for subsequent research. For instance, what will account for the observed differences in function of intonation and lexicosyntax? Lexicosyntactic information is symbolic, conventionally coded, and hierarchically structured. By contrast, features of the speech signal such as intonational contours (in Dutch at least), pauses, and effects of their placement are iconic and indexical, and therefore more likely to be language-independent (see e.g. Ohala 1983 for pitch). There is thus a strong asymmetry in the informational roles of lexicosyntax and intonation: symbolic information strongly determines the interpretation of associated nonsymbolic information.[13] And while the functions of intonation display rich variety, this variety is not infinite. Our study is a case in point: intonation does not give Dutch listeners what they need in order to predict when a current speaker's utterance is going to end. With its high degree of formal flexibility, signaling a wide variety of expressive functions, it may be that for the very task of end-of-turn projection, intonation just cannot compete with the greater restrictiveness, and therefore greater predictiveness, of syntax. While many people seem attached to the idea that it should play a role in turn organization, we think it is only to be expected that there will be limits to the range and number of functions of intonation.

A second line for further research concerns the precise mechanism by which lexicosyntactic structure signals the timing of turn completion. What is it about Dutch lexicosyntax that enables listeners to project turn completion so accurately in free conversation (see Fig. 1, above) as well as in our experimental replication of the same task (Fig. 2)? One avenue is to investigate languages whose grammatical structures differ dramatically

[13] This is amply illustrated in research on co-speech hand gestures, which are often impossible to interpret without the speech that accompanies them, but effortless to interpret when they are accompanied by speech (McNeill 1992, Kendon 2004).

from Dutch, especially where conventional syntactic structures provide less constraining information, for example where there is different and/or more variable word order, or widespread ellipsis. It may turn out that lexicosyntactic structure is universally a better source for end-of-turn projection. To find out why this might be so, one needs to carefully consider the informational properties of unfolding syntactic sequences and the complex relationship between the trajectory of information supply through the course of a syntactic sequence and the available degrees of expressive freedom at any given point. There are conceptual payoffs from this line of inquiry. Just posing the question of how the emerging yet unfinished structure of an utterance can allow listeners to project its completion point forces us to think of lexicosyntax as a temporally unfolding structure which displays (often ambiguous) information about itself through the course of its development. This truly temporal view of lexicosyntax is alien to most corners of descriptive/typological linguistics,[14] but it is de rigueur in the two rather disparate empirical approaches to language that we bring together in this study: experimental psycholinguistics (e.g. Vosse & Kempen 2000) and conversation analysis (e.g. Goodwin 1979, 2006).

The turn-taking problem forces us to think about syntax and cognition in new ways. Grammar is not just a means for semantic representation of predicate-argument relations (or different construals thereof; Langacker 1987, Wierzbicka 1988, Lambrecht 1994, Goldberg 1995). Syntactic structure is also an inherently temporal resource for listeners to chart out the course of a speaker's expression, and to plan their own speech accordingly. In anticipation of this, a speaker may exploit available lexicosyntactic options to actively foreshadow structure, thereby manipulating the unfolding course of the interaction itself, and literally controlling the interlocutor's processes of cognition and action. Thus, the virtues of studying conversation are not just that we are seeing language in its natural home (Thompson & Hopper 2001:27), but that we are seeing how it may be structurally DESIGNED FOR strategic deployment in social interaction (Sacks et al. 1974). These considerations add to the range of factors that may help us understand why languages are structured in just the ways they are.

In closing, we turn to methodological implications. The mode of research that has made significant headway in the domain of conversation is the observational analysis of spontaneously occurring conversation, the tradition known as conversation analysis (Sacks et al. 1974, Levinson 1983, Heritage 1984). A key methodological insight is that much evidence for underlying structure can be found in the observable details of natural conversation itself. Each successive contribution reveals participants' own analysis of the conversational moves being made. But for this study, we needed a different source of evidence. In natural conversation, each occasion of turn transition is unique, supplying one occasion of response, and thus one and only one 'indigenous analysis' of any given utterance. What we needed was an entire distribution of such analyses by multiple speakers in response to a single stimulus. In isolating, varying, and testing competing hypotheses, we endeavored to meet the challenge of balancing experimental control and ecological validity. With stimuli culled from real interactions, and with a real-time projection task, we simulated the psychological conditions of

[14] Notable exceptions include 'emergentist' views such as Hopper 1998 and Ford et al. 2003. However, since these approaches explicitly reject the idea that syntax is abstractly represented, they offer no account for a listener's ability to recognize and project larger oncoming structures on the basis of initial segments of a currently emerging utterance. Nor, for the same reason, can they accommodate the idea that a speaker may strategically deploy a given structure, designed so as to be recognized to be following a certain course.

projection in the wild. The results show that our measure was valid. Just as direct inspection of conversational data will answer questions that experimental methods cannot, our design yielded answers that could not be supplied by direct inspection of conversational materials. The study demonstrates that conversation as an organized structural and psychological domain is amenable to experimental investigation. We see this kind of work not as an alternative to conversation-analytic or descriptive linguistic work, but as a necessary counterpart, where the insights flow in both directions. May the result signal a new trend in cross-disciplinary research on a topic at the very heart of language: the structure of conversation.

REFERENCES

ANDERSON, ANNE H.; MILES BADER; ELLEN GURMAN BARD; ELIZABETH A. BOYLE; GWYNETH DOHERTY-SNEDDON; SIMON C. GARROD; STEPHEN ISARD; JACQUELINE KOWTO; JAN MCALLISTER; CATHERINE SOTILLO; HENRY S. THOMPSON; and REGINA WEINERT. 1991. The HCRC Map Task corpus. *Language and Speech* 34.351–66.

AUER, PETER. 1996. On the prosody and syntax of turn-continuations. In Couper-Kuhlen & Selting 1996b, 57–100.

BEATTIE, GEOFFREY; ANNE CUTLER; and MARK PEARSON. 1982. Why is Mrs. Thatcher interrupted so often? *Nature* 300.744–47.

BOERSMA, PAUL, and DAVID WEENINK. 2004. Praat: Doing phonetics by computer (Version 4.2). Amsterdam: University of Amsterdam. Online: http://www.fon.hum.uva.nl/praat/.

BOLDEN, GALINA B. 2003. Multiple modalities in collaborative turn sequences. *Gesture* 3.187–212.

BROWN, PENELOPE, and STEPHEN C. LEVINSON. 1978. *Politeness: Some universals in language usage*. Cambridge: Cambridge University Press.

CASPERS, JOHANNEKE. 2003. Local speech melody as a limiting factor in the turn-taking system in Dutch. *Journal of Phonetics* 31.251–76.

CLARK, HERBERT H., and JEAN E. FOX TREE. 2002. Using *uh* and *um* in spontaneous speaking. *Cognition* 84.73–111.

COUPER-KUHLEN, ELIZABETH, and MARGARET SELTING. 1996a. Towards an interactional perspective on prosody and a prosodic perspective on interaction. In Couper-Kuhlen & Selting 1996b, 11–56.

COUPER-KUHLEN, ELIZABETH, and MARGARET SELTING (eds.) 1996b. *Prosody in conversation*. Cambridge: Cambridge University Press.

CUTLER, ANNE, and MARK PEARSON. 1986. On the analysis of prosodic turn-taking cues. *Intonation and discourse*, ed. by Catherine Johns-Lewis, 139–55. London: Croom Helm.

DUNCAN, STARKEY. 1974. On the structure of speaker-auditor interaction during speaking turns. *Language in Society* 3.161–80.

DUNCAN, STARKEY, and DONALD W. FISKE. 1977. *Face-to-face interaction: Research, methods and theory*. Hillsdale, NJ: Lawrence Erlbaum.

FORD, CECILIA E., and SANDRA A. THOMPSON. 1996. Interactional units in conversation: Syntactic, intonational, and pragmatic resources for the management of turns. *Interaction and grammar*, ed. by Emanuel A. Schegloff and Sandra A. Thompson, 135–84. Cambridge: Cambridge University Press.

FORD, CECILIA E.; BARBARA FOX; and SANDRA A. THOMPSON. 2003. Social interaction and grammar. *The new psychology of language: Cognitive and functional approaches to language structure*, ed. by Michael Tomasello, 119–44. Mahwah, NJ: Lawrence Erlbaum.

GOLDBERG, ADELE E. 1995. *Constructions: A construction grammar approach to argument structure*. Chicago: University of Chicago Press.

GOODWIN, CHARLES. 1979. The interactive construction of a sentence in a natural conversation. *Everyday language: Studies in ethnomethodology*, ed. by George Psathas, 97–121. New York: Irvington.

GOODWIN, CHARLES. 1981. *Conversational organization: Interaction between speakers and hearers*. New York: Academic Press.

GOODWIN, CHARLES. 2006. Human sociality as mutual orientation in a rich interactive environment: Multimodal utterances and pointing in aphasia. *Roots of human sociality: Culture, cognition and interaction*, ed. by N. J. Enfield and Stephen C. Levinson, 97–125. London: Berg.

GROSJEAN, FRANÇOIS. 1980. Spoken word recognition processes and the gating paradigm. *Perception & Psychophysics* 28.267–83.

GROSJEAN, FRANÇOIS, and CENDRINE HIRT. 1996. Using prosody to predict the end of sentences in English and French: Normal and brain-damaged subjects. *Language and Cognitive Processes* 11.107–34.

GUSSENHOVEN, CARLOS. 2005. Transcription of Dutch intonation. *Prosodic typology: The phonology of intonation and phrasing*, ed. by Sun-Ah Jun, 118–45. Oxford: Oxford University Press.

HAYASHI, MAKOTO. 2005. Joint turn construction through language and the body: Notes on embodiment in coordinated participation in situated activities. *Semiotica* 156.21–53.

HERITAGE, JOHN. 1984. *Garfinkel and ethnomethodology*. Cambridge: Polity Press.

HOPPER, PAUL. 1998. Emergent grammar. *The new psychology of language: Cognitive and functional approaches to language structure*, ed. by Michael Tomasello, 155–75. Mahwah, NJ: Lawrence Erlbaum.

JEFFERSON, GAIL. 1973. A case of precision timing in ordinary conversation: Overlapped tag-positioned address terms in closing sequences. *Semiotica* 9.47–96.

KENDON, ADAM. 1967. Some functions of gaze-direction in social interaction. *Acta Psychologica* 26.22–63.

KENDON, ADAM. 2004. *Gesture: Visible action as utterance*. Cambridge: Cambridge University Press.

LAMBRECHT, KNUD. 1994. *Information structure and sentence form: Topic, focus and the mental representations of discourse referents/grammatical relations*. Cambridge: Cambridge University Press.

LANGACKER, RONALD W. 1987. *Foundations of cognitive grammar, vol. 1: Theoretical prerequisites*. Stanford, CA: Stanford University Press.

LEVINSON, STEPHEN C. 1983. *Pragmatics*. Cambridge: Cambridge University Press.

LOCAL, JOHN; JOHN KELLY; and BILL WELLS. 1986. Towards a phonology of conversation: Turn-taking in Tyneside English. *Journal of Linguistics* 22.411–37.

LOCAL, JOHN; BILL WELLS; and MARK SEBBA. 1985. Phonetic aspects of turn delimination in London Jamaican. *Journal of Pragmatics* 9.309–30.

MAYNARD, SENKO K. 1989. *Japanese conversation: Self contextualization through structure and interactional management.* Norwood, NJ: Ablex.

MCNEILL, DAVID. 1992. *Hand and mind*. Chicago: Chicago University Press.

OHALA, JOHN J. 1983. The origin of sound patterns in vocal tract constraints. *The production of speech*, ed. by Peter F. MacNeilage, 189–216. New York: Springer.

REISMAN, KARL. 1974. Contrapuntal conversations in an Antiguan village. *Language in its social setting*, ed. by Richard Bauman and Joel Sherzer, 110–24. Cambridge: Cambridge University Press.

ROSSANO, FEDERICO. 2005. *On sustaining vs. withdrawing gaze in face-to-face interaction*. Paper presented at 91st annual convention of the National Communication Association, Boston, MA, November 2005.

SACKS, HARVEY; EMANUEL A. SCHEGLOFF; and GAIL JEFFERSON. 1974. A simplest systematics for the organization of turn-taking for conversation. *Language* 50.696–735.

SCHAFFER, DEBORAH. 1983. The role of intonation as a cue to turn taking in conversation. *Journal of Phonetics* 11.243–57.

SCHEGLOFF, EMANUEL A. 1982. Discourse as an interactional achievement: Some uses of 'uh huh' and other things that come between sentences. *Analyzing discourse: Text and talk*, ed. by Deborah Tannen, 71–93. Washington, DC: Georgetown University Press.

SCHEGLOFF, EMANUEL A. 1996. Turn organization: One intersection of grammar and interaction. *Interaction and grammar*, ed. by Elinor Ochs, Emanuel A. Schegloff, and Sandra A. Thompson, 52–133. Cambridge: Cambridge University Press.

SCHEGLOFF, EMANUEL A. 2000. Overlapping talk and the organization of turn-taking for conversation. *Language in Society* 29.1–63.

SHANNON, CLAUDE E. 1948. A mathematical theory of communication. *Bell System Technical Journal* 27.379–423, 623–56.

SIDNELL, JACK. 2001. Conversational turn-taking in a Caribbean English Creole. *Journal of Pragmatics* 33.1263–90.
TANNEN, DEBORAH. 1985. Cross cultural communication. *Handbook of discourse analysis*, ed. by Teun A. Van Dijk, 203–15. London: Academic Press.
THOMPSON, SANDRA A., and PAUL J. HOPPER. 2001. Transitivity, clause structure and argument structure: Evidence from conversation. *Frequency and the emergence of linguistic structure*, ed. by Joan L. Bybee and Paul J. Hopper, 27–60. Amsterdam: John Benjamins.
VOSSE, THEO, and GERARD KEMPEN. 2000. Syntactic structure assembly in human parsing: A computational model based on competitive inhibition and a lexicalist grammar. *Cognition* 75.105–43.
WELLS, BILL, and SUE PEPPÉ. 1996. Ending up in Ulster: Prosody and turn-taking in English dialects. In Couper-Kuhlen & Selting 1996b, 101–30.
WIERZBICKA, ANNA. 1988. *The semantics of grammar.* Amsterdam: John Benjamins.
YNGVE, VICTOR H. 1970. On getting a word in edgewise. *Chicago Linguistic Society* 6.567–78.

Max Planck Institute for Psycholinguistics
P.O. Box 310
NL-6500 AH Nijmegen
The Netherlands

[Received 2 May 2005;
accepted 21 October 2005]