Timing in conversation: The anticipation of turn endings

Lilla Magyari
Max Planck Institute for Psycholinguistics
P.O. Box 310, 6500 AH Nijmegen, The Netherlands
Lilla.Magyari@mpi.nl

Jan Peter de Ruiter
Max Planck Institute for Psycholinguistics
P.O. Box 310, 6500 AH Nijmegen, The Netherlands
JanPeter.deRuiter@mpi.nl

Abstract

We examined how communicators can switch between speaker and listener role with such accurate timing. During conversations, the majority of role transitions happens with a gap or overlap of only a few hundred milliseconds. This suggests that listeners can predict when the turn of the current speaker is going to end. Our hypothesis is that listeners know *when* a turn ends because they know *how* it ends. Anticipating the last words of a turn can help the next speaker in predicting when the turn will end, and also in anticipating the content of the turn, so that an appropriate response can be prepared in advance. We used the stimuli material of an earlier experiment (De Ruiter, Mitterer & Enfield, 2006), in which subjects were listening to turns from natural conversations and had to press a button exactly when the turn they were listening to ended. In the present experiment, we investigated if the subjects can complete those turns when only an initial fragment of the turn is presented to them. We found that the subjects made better predictions about the last words of those turns that had more accurate responses in the earlier button press experiment.

1 Introduction

During conversations, a turn not only has to be relevant to the course of social interaction, but it also has to be appropriately timed. Sacks, Schegloff and Jefferson (1974) assume that transitions from one speaker to the next are accurately timed, so that gaps (silences between turns) and overlaps (i.e. when the interlocutors speak at the same time) are small. It is a normative rule of conversation that requires the participants to respond to the current speaker as soon as he/she has finished. When there are departures from this rule, the gaps or overlaps are interpreted communicatively. For example, a short silence before a response can indicate that the response is a disagreement when disagreement is a dispreferred action (Pomerantz, 1984).

Sacks et al.'s normative rule has been recently supported by measurement of floor transfer offset (FTO) in a data-set from Dutch two-party telephone conversations (De Ruiter et al., 2006). The FTO is defined as the difference between the time that a turn starts and the moment the previous turn ends. In the Dutch conversations, 45% of all speaker transitions had an FTO of between -250 and +250 ms, and 85% of them were between -750 and 750 ms. (Negative values indicate an overlap between the consecutive turns, positive values indicate a gap.) The FTO values were centered around 0. This pattern supports Sacks et al.'s (1974) assumption that gaps and overlaps are small. However, it also raises the question of how these accurately timed transitions are possible.

Such accurate temporal alignment of conversational turns suggests that a potential next speaker can anticipate the moment when the current turn is going to end. If the next speaker detects the end of the current turn (but she does not anticipates it), she will have a little delay before her turn because preparation for articulation requires some time. However, many turn transitions happen without temporal gaps.

Sacks et al. have already assumed that the potential next speakers can plan to align their turn accurately in time only if they are able to accurately predict the end of the current speakers turn. However, they left open the question of exactly how the anticipation of end of turns is carried out.

Many sources of information (semantic, syntactic, pragmatic and prosodic) have been proposed to be used in the prediction of turn endings. The few experimental studies which

have investigated this issue, mainly concentrated on the role of intonation in end-of-turn predictions. Groesjean and Hirt's study (1996) investigated if people can use prosodic information to predict end of French and English sentences. Subjects were listening to sentences that were presented in segments of increasing duration. They had to guess with how many words the fragments would continue. The sentences of which the initial fragment was presented to the subjects were either short or they were expanded by optional noun-phrases. The subjects had to guess using a multiple choice response task if the presented fragment was part of a short sentence or an expanded, longer sentence. The predictions did not improve with increasing duration of the fragments (sentence beginnings). Only when the first potentially last word was presented (i.e. the first point in the sentence where the sentence could end if it would be a short sentence) could the subjects predict if the sentence would be finished after the potentially last word or it would continue with 3 or 6 more words. According to Grosjean and Hirt the results indicate that in English prosodic information is made available for the prediction of sentence length only when the semantic and syntactic information can not help. Their similar experiment on French showed that subjects could tell if a sentence has ended or not. But they could not predict with how many words the sentences (3, 6 or 9 more words) would continue.

Grosjean and Hirt's study used recordings of sentences read aloud. The prosodic pattern may differ from the prosody occurring in natural conversations. Therefore, it is questionable how their results can be generalized to account for processing of spontaneous speech.

De Ruiter et al. (2006) investigated the contribution of the lexico-syntactic content and intonation in end-of-turn predictions. They manipulated recordings of natural conversations. Subjects listened to individual turns taken out from Dutch telephone conversations. They were asked to press a button exactly at the moment the turn ended. The duration between the end of the turn and the button-presses (called *bias*) was measured. In the different experimental conditions, the turns were presented naturally (as recorded) or a modified version was played. In one of the conditions, the intonational contour was removed, in another condition the lexico-syntactic content was removed by applying low-pass filtering. When subjects were listening to the original turns, their button-presses coincided with the turn-ends accurately; the distribution of the button-presses was similar to the distribution of FTO values for the same turns in the original conversations. There was no change in accuracy when the intonational contour was removed, but the performance got worse when the words could not be understood (note that the intonational information was still present in those stimuli). De Ruiter et al. concluded that the intonational contour is neither necessary nor sufficient for the prediction of turn-ends. These results suggest the lexico-syntactic information plays a major role in timing of turns.

Listeners have to perform many simultaneous tasks before they start their turn. They have to perceive and comprehend the current turn, and also formulate and time their subsequent utterance appropriately. The fine temporal alignment of conversational turns shows that these tasks have to be done simultaneously. Response preparation has to start before the previous turn ends in order to avoid gaps. Response preparation, however, can be initiated only if the speaker knows roughly what to respond. Therefore, the next speaker has to anticipate not only the end of the turns but also their content. When the last words of a turn can be anticipated they give information about the content and about the duration in advance. Therefore, we hypothesize that lexico-syntactic information helps in the prediction of the time when a turn will end through the anticipation of the last words of a turn. In other words: People know *when* a turn ends by knowing *how* it ends.

In order to test this hypothesis we conducted an experiment using the experimental stimuli of De Ruiter et al.'s study. Our prediction was that the more accurate the button-presses to the end of a given turn were in the earlier experiment, the more accurately the last words of that turn can be predicted. Therefore, we examined if there was any correlation between the accuracy of button presses in the earlier experiment and the off-line prediction of last words of the turn in a gating study. The end of selected turns were cut off at several points and fragments or the entire turn were presented to subjects who then had to guess how the turn would continue.

2 Methods

2.1 Participants

Fifty native speakers of Dutch (forty-two women and eight men, aged between eighteen and

twenty-nine) participated in the experiment. The data of one subject was excluded because the results showed that he did not understand the task correctly. The subjects were paid for their participation.

2.2 Stimulus material

The experimental materials were selected from stimuli used by De Ruiter et al. These stimuli were turns from natural conversations in Dutch. In the De Ruiter et al. experiment, it had been measured for each turn how accurately subjects could predict the end of turns by button-press. The temporal offset (bias) between the end of the turn and the button-presses was measured. The averaged bias of a turn indicates how accurately subjects could on average predict the time point of the end of that turn. A turn with a highly positive bias means that subjects pressed the button too late. A low bias (small positive value or with a small negative value) shows that subjects pressed the button on time or a bit earlier, just before the turn ended.

For the purposes of the present study, turns with high and low biases from the De Ruiter et al. study were selected. It was observed that turns with longer duration tend to have a lower bias. In order to avoid effects caused by the duration of the turns, ten turn-pairs were selected, where both members of the pairs had the same duration. The members of each pair were from different conversations produced by different speakers. The members of each pair had the same duration (max. difference between the members of the pairs was 16 ms), but they differed in their average bias. One of the members of every pair had higher average bias (between 237 and 123 ms), while the other member had a lower average bias (between -18 and 122 ms) relative to the other member of the pair. The durations of the 10 stimuli pairs were varying between 1.13 s and 2.05 s.

For each turn pair, four versions were made by cutting off the speech at four different temporal locations. The cut-off locations within each pair were at same points in time measured from the end of the recordings, but they were different across stimuli pairs. The cut-off locations were determined in a pair according to the boundaries of the two last words of each of the pairs. Each stimulus was cut at four points which were just at word boundaries at one of the members of a stimuli pair. The cut-off location varied across the pairs (the first points were on average at 0.76s from the end, the second points at 0.52s; the third points at 0.40 s; the fourth points at 0.25 s). Table 1. shows an example of gating points of one of the turn pairs that was used in the experiment. Turn A and B have almost the same duration (1.78 and 1.79 s), while A is a low bias turn (40 ms) and B is a high bias turn (226 ms). The vertical lines shows the points where both turns were cut in order to create the fragments. The vertical lines that are aligned with each other between the two turns indicate that the cut-off was made at the same points in time measured from the end of the recordings.

A.
maar dat hoor ik wel via |de|
mi|crof |oon
B.
ja maar daar moeten we maar een keer met| |
zijn|allen| heen

Table 1. Example of a turn-pair and the cut-off locations (shown by the vertical lines in the text)

2.3 Experimental design

Subjects were randomly assigned to one of five experimental lists. The stimuli in the lists were presented in random order to each subject. Their task was to type in if the presented segment constituted a complete turn. If the subjects decided that the turn was not complete, they were asked to guess and type in how they thought it would continue. If they did not have any guess about the continuation, they were asked to guess with how many words the turn would continue. They had to make a forced choice between A. one word, B. two words, or C. three or more words. Subjects were also asked how certain they were of their responses on a four point scale.

2.4 Procedure

The subjects were requested to sit in front of a computer screen and a keyboard with headphones. The instructions were visually presented on the screen. Before each stimulus a sentence was presented on the screen in Dutch, saying: "When you press the space bar you can listen to the next sound fragment two times.". 500 ms after pressing the space bar, a stimulus was presented two times, with a 1500 ms pause between the two presentations. After the stimulus presentation, the subjects saw a prompt (>:) on the screen where they had to type their guess

about the continuation of the fragment. If they thought the turn that they were listening to was complete, they had to type: '.'. If they did not have any guess about the continuation, but they did not think that the turn had finished, they were asked to type a '-'. After reading the instructions, the participants did a training session during which four stimuli were presented that were not part of the experimental list. After the training session, and possibly providing verbal clarifications, the experimenter left the room and the participants could continue the experiment alone.

2.5 Data-coding

Two variables with categories were created based on the responses. The variable PREDEND (prediction of the rest of the turn) was 0 when the continuation of the turn was entirely correct. It was 0 also if it was indicated correctly that the turn has ended. PREDEND could get 0 only if the guess was entirely correct regardless how many words had to be guessed. PREDEND was 1 when it was incorrect: when different words were used, when the end was indicated wrongly or when the participants did not have any guess.

PREDNUM (prediction of the number of words) variable had three categories: 1, when the predicted number of words was the same compared to the original version of the sentence even if the words were not the same, or when the participant did not have any idea about the continuation (but the prediction of the number of words in the continuation was correct); 2, when the predicted number of words was less than the number of words in the turn, and 3, when more words were predicted.

Responses which were not clear (e.g. words that do not exist) were excluded from the analysis. Only 3% of the data points were excluded.

3 Results

3. 1 Statistical analysis

The results were analyzed using a generalized linear mixed effects model (GLMM) (Baayen, 2008, Pinherio & Bates, 2000). We used this statistical analysis because of two main reasons. On one hand, it has been shown that mixed-effects models provide a better method for statistical analysis with repeated measurement data (see for example, Baayen, Davidson & Bates, 2008). Among other advantages, the mixed effect regression model can simultaneously handle all factors that potentially can contribute to explaining the variance in the data. The model can include random effects, such as variations caused by individual differences among the subjects and variations caused by differences in the properties of the items. It is also possible to fit the model to unbalanced data.

GLMM also has many advantages over the widely used repeated-measurement analysis of variance (ANOVA) for categorical datasets. In this experiment, the proportions of correct and non-correct responses were analyzed that do not follow normal distribution that can be problematic for the ANOVA analysis. When ANOVA is used for categorical outcomes, it can yield spurious results that GLMM can avoid (Jaeger, 2008).

3.2 Recognition of turn-ends

Figure 1 shows the percentage of correct responses when subjects were listening to the entire turn. The responses are highly accurate. 96% of the participants give correct responses at high-bias turns, and 90% of the participants at low bias turns.

The PREDEND variable was binary (correct or not correct), therefore a binomial distribution was specified for the model.

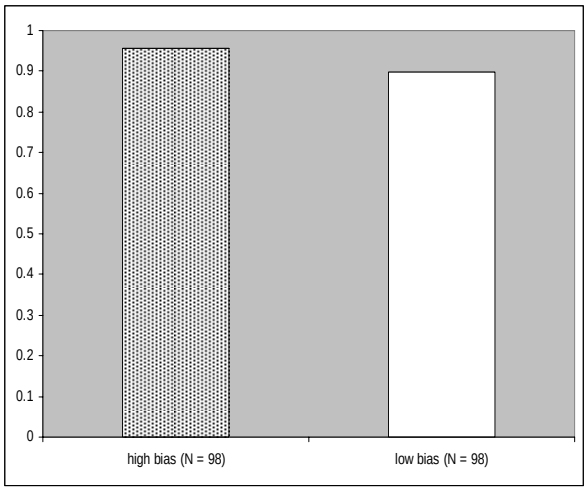

Figure 1. The proportion of correct responses ('the turn has ended') when the entire turn was presented

The linear model had Bias (if the turn belonged to the high or low bias turns) as a fixed effect, and Subjects and Utterance-pairs as

random effects. The GLMM analysis did not show any effect of Bias ($z = 1.498$, $p>0.1$, $N = 196$).

3.3 Prediction of the continuations

Figure 2 shows the proportion of the correct continuations at each cut-off location for both turn types. From the first cut-off location (I) to the fourth (IV) increasing proportion of the turns were presented to the subjects. The proportion of correct answers is increasing as the presented fragments get longer.

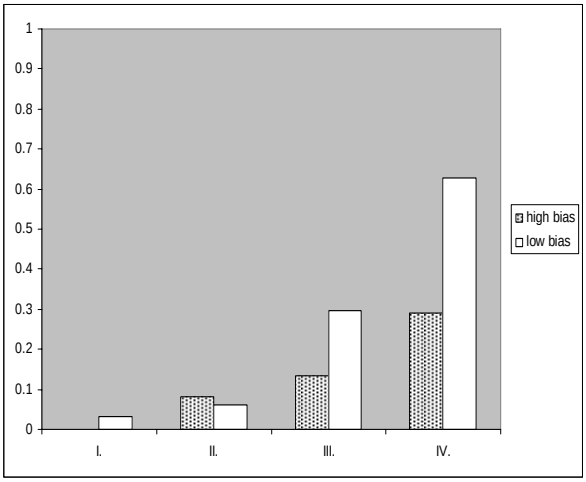

Figure 2. Proportion of the correct continuations at each cut-off locations. The x-axis shows the proportion (between 0 and 1), the y-axis shows the cut-off locations (from I. to IV. the duration of fragments from each turn are increasing). The white columns show the proportion of correct continuations when a fragment from a low bias turn was presented, the grey columns show the proportion of correct answers that belong to the high bias turns.

However, it is possible that differences between the two turn types may arise from the properties of the stimuli material. Some fragments were cut so close to the end of the last word that it sounded as the end. Therefore, the correct response was that the turn has ended and not a free guess about the continuation. It is probably easier to decide if a turn continues or not than it is to predict its continuation. Therefore, those turns where despite of the cut off, there was no more reliable auditory information coming, were excluded from our analysis (11%). Figure 3 shows the proportion of the correct responses at the four consecutive cut-off locations after the exclusion.

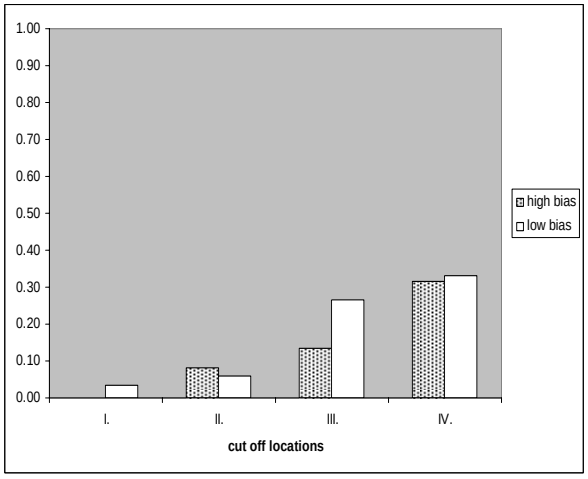

Figure 3. Proportion of correct continuations after excluding some of the fragments. For explanation of the figure, see Figure 2.

The differences between the turn types got reduced at the last cut-off location (IV). Table 2 shows the number of items at each cut-off location for both turn types and the proportion of the correct responses.

bias	I.	II.	III.	IV.
high	0.00 (N=96)	0.08 (N=98)	0.13 (N=98)	0.32 (N=68)
low	0.03 (N=98)	0.06 (N=98)	0.27 (N=78)	0.33 (N=60)

Table 2. The proportion of the correct continuations (1 = all are correct) and the number of items at each cut-off locations for both turn-types

At the last two cut off locations (III and IV) at 20% and at 31.5% of the cases subject were able to guess the correct continuations.

The GLMM had Bias and Cut-off location as fixed effects, and Subjects and Utterance-pairs as random effects. Table 3. shows the β-coefficients of the fixed effects in the model.

Predictor	Coeff	SE	z value	p
Intercept	6.578	0.774	8.503	<0.001
Cutoff	-1.318	0.166	-7.962	<0.001
Bias	-0.674	0.302	-2.235	<0.05

Table 3. The summary of the fixed effects in the GLMM of correct continuations at the four cut-off locations. (Coeff = Coefficient)

Both the cut-off locations (z -7.962, $p<0.001$, $N=694$) and the bias ($z=-2.235$, $p<0.05$, $N=694$) had a significant effect on the correct responses. It is possible, however, that low bias turns were

easier to complete because they always ended in a longer word than high bias turns. It means for example, that at the last cut-off location (IV) the fragments ended during the last word at the low bias turns, while the fragments ended before the last word at the high bias turns. In this case, maybe it is easier to recognize a word that was partially played than to guess for a not-heard at all word. In order to explore if this explanation is valid, the proportion of fragments that ended during the last word and before the last word at the fourth cut-off location was calculated for both types of turns (Figure 4).

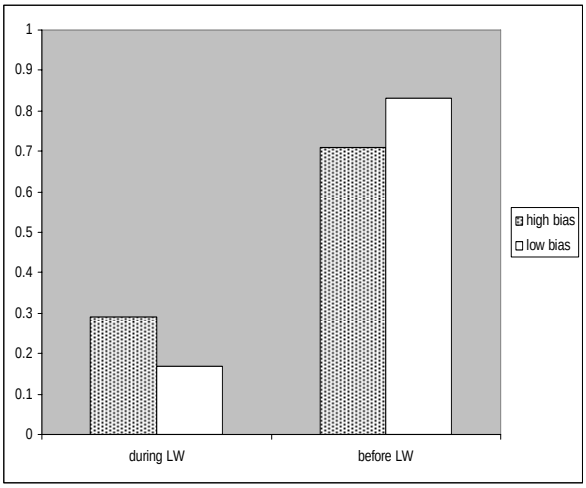

Figure 4. Proportion of stimuli that was cut during the last word (LW) or before the last word at the last cut-off location

The cut-off locations occurred during the last word in a smaller percentage at the low bias turns than at the high bias turns. The GLMM shows a significant effect ($z=2.5375$, $p<0.05$, $N=128$) of the position of the cut-off location (WB, during or before the last word) on the correct continuations. However, the direction of the effect is in the opposite direction. The guesses get better when the cut-off location is before the last word and not during it. Therefore, the observed differences between the turn types can not have been caused by the earlier recognition of the last words.

3.4 Prediction of the number of words

Our question was also if the difference between the high and low bias turns is not only caused by anticipation of the correct turn endings but also by the prediction of the correct number of words, irrespective of their form or meaning. We examined if there is a correlation between the turn types and the expectations about the length of the turn even if the continuations are wrong. Therefore, the number of the predicted words (PREDNUM) was analyzed for the cases where the prediction of the actual continuation was not correct. We again excluded those turns that has already finished at the two last cut-off locations.

Figure 5 shows the proportion of correct guesses about the number of words. We expected to find a higher proportion of correct responses at the low bias turns because that could lead to accurate button-presses. But Figure 5 shows the opposite direction of the differences.

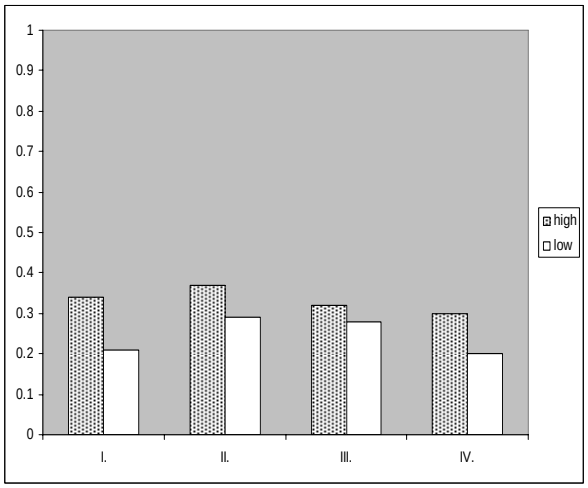

Figure 5. Proportion of correct guesses of the number of the coming words among the wrong continuations. For more explanation of the figure, see Figure 2.

Bias and the Cut-off locations were included as fixed effects, while Subjects and Utterance-pairs were included as random effects in the linear mixed effects regression analysis of the correct number of words estimates among the wrong guesses. The analysis showed a main effect of Bias ($z=2.56$, $p<0.05$, $N=601$) (Table 4).

Predictor	Coefficient	SE	z value	p
Intercept	0.611	0.265	2.307	$p<0.05$
Bias	0.476	0.186	2.56	$p<0.05$
Cutoff	0.035	0.088	0.399	$p>0.05$

Table 4. The summary of the fixed effects in the GLMM of correct estimates of the number of words among the wrong guesses

Figure 6 shows the proportion of the responses that predicted less number of words than the number of words that were still coming but not played.

The GLMM analysis (Table 5) showed that there is a significant effect of the Cut-off locations ($z=5.029$, $p<0.001$, $N=601$), and that

there is an interaction between Bias and Cut-off locations (z=-5.071, p<0.001, N=601). The difference between turn types in the less number of words predictions are increasing towards the end of the turn. The low bias turns tend to have a higher proportion of less number of words guesses.

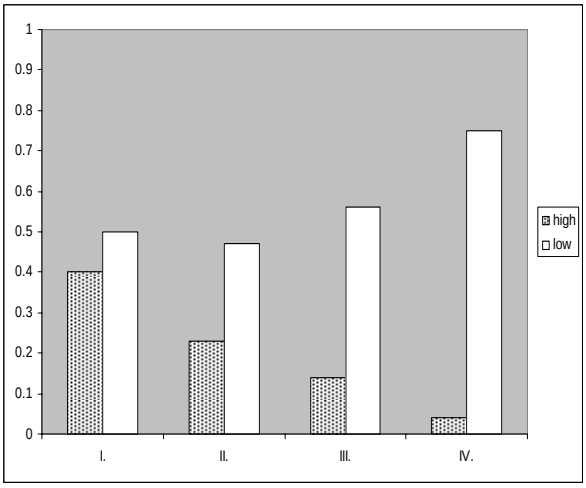

Figure 6. Proportion of guesses predicting less number of coming words among the wrong continuations. For more explanation of the figure, see Figure 2.

Predictor	Coeff	SE	z value	p
Intercept	-0.394	0.402	-0.98	p>0.05
Bias	0.588	0.445	1.322	p>0.05
Cutoff	0.828	0.165	5.029	p<0.001
Interaction: Bias&Cutoff	-1.06	0.209	-5.071	p<0.001

Table 5. The summary of the fixed effects in the GLMM of estimates of less number of words among the wrong guesses. (Coeff = Coefficient)

Figure 7 shows the proportion of the responses that predicted more number of words than the number of words that were still coming but not presented. The regression analysis showed an effect of Bias (z=-2.154, p<0.05, N=601) and Cut-off locations (z=-4.895, p<0.001, N=601), and also their interaction (z=4.836, p<0.001, N=601). More number of words were predicted at the high bias turns than at the low bias turns (Table 6).

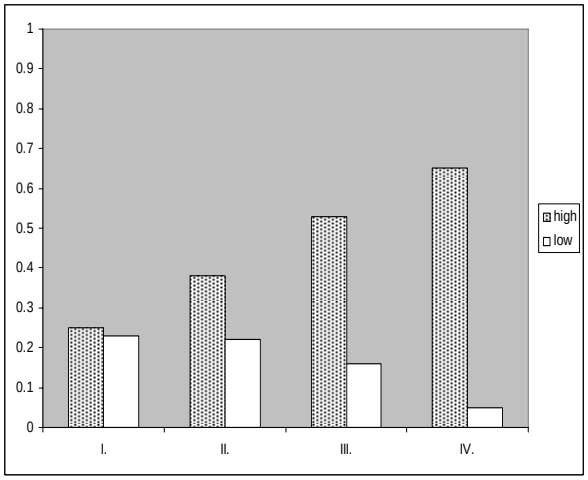

Figure 7. Proportion of guesses predicting more number of coming words among the wrong continuations. For more explanation of the figure, see Figure 2.

Predictor	Coeff	SE	z value	p
Intercept	1.655	0.316	5.242	p<0.001
Bias	-0.987	0.458	-2.154	p<0.05
Cutoff	-0.582	0.119	-4.895	p<0.001
Interaction: Bias&Cutoff	0.983	0.203	4.836	p<0.001

Table 6. The summary of the fixed effects in the GLMM of estimates of more number of words among the wrong guesses. (Coeff = Coefficient)

In order to see how early the differences in the number of words predictions are present, an additional analysis was done for the first two cut-off locations (I and II). Fragments with these cut-off locations ended before the last word in all cases. A mixed effect regression model was fitted for the cut-off location I and II separately with Bias as main effect, and Utterance pairs and Subjects as random effects. Bias had an effect at both cut-off locations when less number of words were predicted: At location I, z=-2.187, p<0.05, N=191 and at location II, z=-3.647, p<0.01, N=182. Bias did not have an effect at the first cut-off location (z=0.298, p>0.05, N=191), but it had an effect at the second cut-off location (z=2.35, p<0.05, N=182) when more number of words were predicted. This means that the subject predicted in a higher proportion less number of words at the low bias turns, and more number of words at the high bias turns by listening to fragments that did not contain the last word of the turns.

4 Discussion

We investigated the hypothesis that people know when a turn ends because they know how it ends. Gating paradigm was used to examine if it is possible to predict the last words of conversational turns. The turns were extracted from natural conversations and the original context was not presented to the subjects. Even so the subjects could guess the not presented or only partially presented last words of the turns correctly in around 20 - 30% of the cases. We found differences also among the turn-types. The continuation of turns whose ends were indicated too late by the button-pressing task in an earlier experiment were less often predicted correctly than the continuations of those turns whose ends were indicated on time (low bias turns). We have shown that these differences between turn-types could not have been caused by earlier or later recognition of the last word of that turn.

We also found that when the continuations were not correct subjects predicted less numbers of words at the low bias turns, and more numbers of words at the high bias turns before the last word of a turn. This shows that probably when the subjects thought that more words were coming, they pressed the button too late in the button-press experiment. These results support the hypothesis that the prediction of turn endings is based on the predictions made about the content and word forms of the turn.

This study emphasis the role of anticipation in order to explain the alignment of turns during conversations. This is in line with studies that show that people use the linguistic context for anticipating the upcoming words (DeLong, Urbach & Kutas, 2005). Pickering and Garrod (2007) argue that comprehenders use the production system for making predictions in order to achieve a faster and easier comprehension. However, investigation of conversational turn alignments shows that it is very likely that preparation of responses occurs in temporal overlap with the comprehension or prediction of the current turn. If the production system is used for predicting the upcoming words, then the same system is used also for preparation of the coming turn. We doubt whether the same system can fulfill two tasks at the same time. We think it is more plausible to assume that the comprehension system is used for anticipation of the content and word forms of the current turn, while the production system is used for preparation of the next turn.

This off-line study has some limitations in generalizing its results to on-line processing. However, the study shows that people can make accurate predictions about the final word forms of turns from natural conversations. The results also suggest that anticipation about the number of words of a turn can explain the accurate performance in turn-end predictions.

References

Katherine A. DeLong, Thomas P. Urbach and Marta Kutas. 2005. Probabilistic word pre-activation during language comprehension inferred from electrical brain activity. *Nature Neuroscience*, 8:1117-1121.

Harald Baayen. 2008. *Analyzing Linguistic data: A practical introduction to statistics.* Cambridge University Press.

Harald Baayen, Doug Davidson and D.M. Bates. 2008. Mixed-effects modeling with crossed random effects for subjects and items. *Journal of Memory and Language*, in press, doi:10.1016/j.jml.2007.12.005.

Francois Grosjean 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.

T. Florian Jaeger. 2008. Categorical data analysis: Away from ANOVAs (transformation or not) and towards logit mixed models. *Journal of Memory and Language*, in press, doi:10.1016/j.jml.2007.11.007.

Martin J. Pickering and Simon Garrod. 2007. Do people use language production to make predictions during comprehension?. *TRENDS in Cognitive Sciences*, 11:105-110.

J. C. Pinherio and D. M. Bates. 2000. *Mixed-Effects Models in S and S-PLUS*. Springer.

Anita Pomerantz. 1984. Agreeing and disagreeing with assessments: some features of preferred/dispreferred turn shapes. In J. M. Atkinson and J. Heritage (Eds.). *Structures of Social Action*. Cambridge, Cambridge University Press, 53-101.

J. P. de Ruiter, Holger Mitterer, and Nick J. Enfield. 2006. Projecting the end of a speaker's turn: A cognitive cornerstone of conversation, *Language*, 82:515-535.

Harvey Sacks, Emanuel A. Schegloff and Gail Jefferson. 1974. A simplest systematics for the organization of turn-taking for conversation. *Language*, 50:696-735.