# Durational aspects of turn-taking in spontaneous face-to-face and telephone dialogues

Louis ten Bosch[1], Nelleke Oostdijk [1], Jan Peter de Ruiter [2]

[1] Dept. of Language and Speech, Nijmegen University, the Netherlands
(l.tenbosch@let.kun.nl)
[2] Max Planck Institute for Psycholinguistics, Nijmegen, the Netherlands

**Abstract.** On the basis of two-speaker spontaneous conversations, it is shown that the distributions of both pauses and speech-overlaps of telephone and face-to-face dialogues have different statistical properties. Pauses in a face-to-face dialogue last up to 4 times longer than pauses in telephone conversations in functionally comparable conditions. There is a high correlation (0.88 or larger) between the average pause duration for the two speakers across face-to-face dialogues and telephone dialogues. The data provided form a first quantitative analysis of the complex turn-taking mechanism evidenced in the dialogues available in the 9-million-word Spoken Dutch Corpus.

## 1 Introduction

Turn-taking in human-human dialogue is a highly complex phenomenon. In order to maintain a smooth dialogue, speakers employ turn-keeping and turn-yielding cues to signal their intention to keep or willingness to yield the turn. Turn-taking in dialogues has received substantial interest during the past decades. Sacks et al. (1974) describe turn-taking as a set of rules adhered to by speakers. In their framework, speaker changes can only happen at specific moments determined by prosodic, pragmatic, syntactic and semantic factors. The smooth alternation of speaker and listener roles in a natural dialogue would then be the result of the aim of the interlocutors to minimize both the duration of speech overlaps and the time lapses between the turns.

More recent studies on turn-taking behavior have shed more light on the complex relation between turn-taking, syntactic and paralinguistic factors (e.g. Ford and Thompson, 1996; Koiso et al., 1998; Caspers, 2001; Selting, 1996). Many of these studies are based on dialogues in special situations, e.g. the Map Task[1]. In the present study, we take up the challenge to investigate turn-taking in spontaneous dialogues. In doing so, we restrict ourselves to a factual description of the durational aspects of turn-taking as observed in these dialogues. However, we first must develop operational definitions of the concepts of 'utterance' and 'turn'.

[1] http://www.hcrc.ed.ac.uk/maptask/maptask-papers.html

## 2 Data, annotations, and analysis method

### 2.1 Data

Our dialogue corpus has been derived from the Corpus Gesproken Nederlands (CGN, Spoken Dutch Corpus, Oostdijk et al., 2002), a 9-million-word corpus comprising a variety of sub-corpora. The corpus has been annotated with many different types of information, including orthography and part-of-speech tags. The orthographic annotation comprises the verbatim transcription, special symbols to mark truncated words or unintelligible speech, and some punctuation (a period signalling the end of an utterance, ellipsis, and a question mark signalling the end of the utterance that is interpreted as a question). The dataset used in the present study consists of 29 face-to-face dialogues and 32 telephone dialogues, and chosen in such a way that a word-level segmentation was available for all selected data (see Oostdijk et al., 2002). Both face-to-face dialogues and telephone dialogues are informal and spontaneous; speakers knew each other and could freely talk about any subject. Each dialogue lasts between 7 and 11 minutes.

### 2.2 Background

For the description of turn-taking, we define a 'turn' as a stretch of speech uttered by one speaker that consists of one or more utterances. An "utterance" is defined as the sequence of words between punctuation marks in the part-of-speech annotation tier. The first issue we address concerns the temporal organization of turns in terms of utterances. The second issue is related to the function of the utterances in a dialogue. Some utterances such as "hm-hm" function as back-channel signals or 'continuers' (Schegloff, 1982), while others carry propositional meaning.

A study by Weilhammer and Rabold (2003) on durational aspects of turn-taking, which was based on task-oriented dialogue data, has shown that the logarithm of the durations of pauses and overlaps can be modeled by a Gaussian distribution. In their analysis, the definition of turn was 'implicitly based' on the Verbmobil transcription conventions (Burger, 1997). Their definition of a turn states that 'a turn starts with the first word in the dialogue or with the first word breaking the silence that follows the previous turn'. Furthermore, 'the silence between two turns of one speaker is always overlaid by an utterance of the [interlocutor]'. The definition of a turn as used in the present study is very similar.

### 2.3 Configurations of turns, pauses and overlaps

Weilhammer and Rabold (2003) give an overview of ten different temporal configurations of turns, describing different possible speech starts by speaker A and speaker B. We have developed a similar scheme for the analysis of the CGN data. Figure 1 shows the various possibilities for the temporal relation of utterances by A and B. The diagram refers to the moment where A has the turn and has finished an

utterance A1. It distinguishes several cases according to the start of an utterance (B in the diagram) by speaker B or a second utterance (A2) by speaker A. The labels on the left denote the code for a specific situation. Of the ten possible cases of turn changes that Weilhammer et al. (2003) distinguish, we have collapsed four (1 and 5b, 2 and 5c) into two categories, since they do not differ with respect to the time relation of turn B relative to turn A1. For other situations we have distinguished more subcategories to obtain a more precise description: Weilhammer et al.'s categories 1 and 5b become b1, b2 or b3 in our system, and categories 2 and 5c become a1, a2, or a3 in ours.

The resulting utterance classification has been used to define turns changes. Turn changes from speaker A to speaker B after A1 occur in the cases a2, a3, a4, b1, b2 and b3, while in the cases c1 and c2, A keeps the turn by uttering A2. The cases d1, a1, and z are mentioned for the sake of completeness only: cases z and a1 are covered by the annotation of the previous utterance of A, and d1 is covered by the annotation of A2.

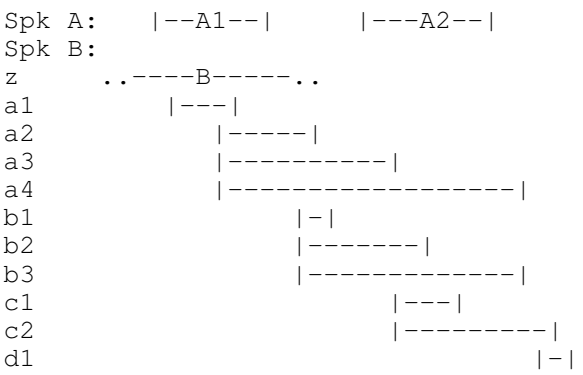


Figure 1. Overview of the different temporal patterns of utterances used to define turns and turn changes. Each horizontal bar refers to an utterance (a sequence of words terminated by a period, an ellipsis, or a question mark). A1 is an utterance by speaker A, B and A2 are utterances by the interlocutor and the next utterance of speaker A, respectively. The labels on the left-hand side denote the various options after A1 finishes. For example, in case a2, B takes over the turn from A before the end of A1. In case c1, B starts after the second utterance of A, thereby defining A's turn to consist of multiple utterances.

On the basis of this classification, durations of pauses and of overlaps have been defined in the following way. The time between the end of A1 and start of B in the cases a2, a3, and a4 counts as overlap, while pauses are the time spans between A1 and B in the cases b1, b2, and b3. Of the remaining categories, only c1 deserves special interest since turns B of type c1 and type b1 allow direct comparison with respect to the content (in the case c1, B is completely overlaid by A2, and likely to be a back-channel; in case b1, B is possibly a back-channel or a propositional utterance). Table I summarizes the definition of pause and overlap between turns of speaker A and B.

Table I. An overview of the utterance annotation, and the corresponding definition for overlap and pause durations.

| Case | duration | description |
|---|---|---|
| z | -- | rest cat |
| a1 | -- | advanced c1 |
| a2 | end(A1)-begin(B) | overlap A1&B |
| a3 | end(A1)-begin(B) | overlap A1&B |
| a4 | end(A1)-begin(B) | overlap A1&B |
| b1 | begin(B)-end(A1) | pause A1-B |
| b2 | begin(B)-end(A1) | pause A1-B |
| b3 | begin(B)-end(A1) | pause A1-B |
| c1 | -- | full overlap |
| c2 | -- | postponed a2 |
| d1 | -- | postponed b1 |

### 2.4 Results

In Table II, absolute and relative frequencies are presented for the various turn types, for the face-to-face (first and third column) and for the telephone situation (second and fourth column). The most salient difference between the two settings is the higher number of overlaps in the telephone dialogues.

Table III shows the difference between face-to-face and telephone dialogues, focussing on the cases that are associated with turns (i.e. a2, a3, a4, b1, b2 and b3). The table shows that the number of 'clean' turns (case b1) is lower in the telephony setting. All other turns relate to an overlap (38 percent for face-to-face, 51 percent for telephony). The partial sum for the cases a2, a3, and a4 shows that the number of turn-takings before the end of an utterance increases from 19 percent to 26 percent.
Figure 2 shows the histograms for the logarithm of the durations of pauses and overlaps (top and bottom panels). For each plot, the x-axis presents the logarithm (base 10) of the durations, while the y-axis presents the number of observations in the corresponding bin. Included are all the cases a2, a3, a4, and b1, b2, and b3.

The histograms of the logarithms of pause durations approximate a Gaussian shape. In comparison to the face-to-face data, the telephony data show a shift towards shorter pause durations. The overlap histograms (lower panels) also appear to have a distribution which approximates a Gaussian shape. Weilhammer et al. (2003) report that the distributions in the VerbMobil data for overlap duration are best modeled as a mix of two-Gaussian distributions, without providing an explanation for the bi-modal character.

With respect to the durations of pauses, telephone dialogues show more much shorter pauses than face-to-face dialogues. There are many more overlaps (all turn types except for the 'clean turn' b1 are more frequent in telephone conversations). We will discuss this finding in more detail in the discussion section below.

Figure 3 illustrates another interesting phenomenon. The figure presents a scatter plot of the average pause duration (measured for each speaker. Each dialogue is represented by a single point in the scatter diagram. The resulting scatter plot shows a high correlation (0.88) between the average pause duration of speaker A and of

speaker B for both telephone as well as the face-to-phase dialogues; furthermore, the variation in average pause durations is much larger (up to a factor 4) in the face-to-face dialogues.

Table II. Absolute and relative counts for face-to-face and telephone data and for each turn category. Gross totals for each column are presented on the first data row. 'Not ann.' refers to 'not annotated': cases where the utterance could not be given a label (in most cases because it was the very last one in the dialogue). The next number is the overall sum of the individual counts of each type. Real turns are a2, a3, a4, b1, b2 and b3. Of these, b1 is the single non-overlapping turn: in the a-case B overlaps with A1, in the case b2 and b3, B overlaps with A2. Categories c1, c2, d1 and z are presented for sake of completeness, a1 is empty by construction.

| Type of turn | Face-to-face (counts) | Telephony (counts) | Face-to-face (perc.) | Telephony (perc.) |
|---|---|---|---|---|
| Total | 8003 | 11583 | 100.0 | 100.0 |
| Not ann. | 83 | 79 | 1.1 | 0.7 |
| Annotated | 7920 | 11504 | 98.9 | 99.3 |
| a2 | 449 | 895 | 5.6 | 7.7 |
| a3 | 138 | 380 | 1.7 | 3.3 |
| a4 | 120 | 319 | 1.5 | 2.7 |
| b1 | 2255 | 2974 | 28.2 | 25.7 |
| b2 | 393 | 882 | 4.9 | 7.6 |
| b3 | 274 | 543 | 3.4 | 4.7 |
| c1 | 506 | 883 | 6.3 | 7.6 |
| c2 | 417 | 773 | 5.2 | 6.7 |
| d1 | 2622 | 2354 | 32.8 | 20.3 |
| z | 746 | 1501 | 9.3 | 13.0 |

Table III. Relative frequencies for the real turns, associated with cases a2, a3, a4 (overlaps) and b1, b2 and b3 (pauses). The normalisation is such that the categories 'a' and 'b' make the total of 100 percent.

| Case | Face-to-face (percentages) | Telephone (percentages) | Partial sums for cases a2-4 and b1-3 |
|---|---|---|---|
| a2 | 12.4 | 14.9 | |
| a3 | 3.8 | 6.3 | |
| a4 | 3.3 | 5.3 | |
| | | | 19.4, 26.6 |
| b1 | 62.1 | 49.6 | |
| b2 | 10.8 | 14.7 | |
| b3 | 7.5 | 9.0 | |
| | | | 80.5, 73.4 |
| total | 100.0 | 100.0 | |

## 2.5 Analysis of turns of type b1 and c1

Until now, we have presented a description of the turn-taking phenomena with emphasis on the temporal aspects. In this section, we further analyze the difference between the 'real' turns by B (case b1, in which A remains silent) and the turns by B that are overlapped by A (case c1). An analysis by hand of these turns led to the observation that utterances from a speaker can be broadly classified into 4 types:

1) back-channels (very short, one to five tokens: *um, mmm, ja, goh zeg, dat zal wel ja*)
2) Failed attempts to take over the turn (usually rather short: e.g. *ik ben uh …, maar da's uh …, hé maar …*)
3) Short propositional utterances that provide some feedback to the previous utterance or turn (content-based, e.g. *grappig [funny], da's wel substantieel*)
4) Longer actual propositional phrases (e.g *heb je 't ook druk gehad?*)

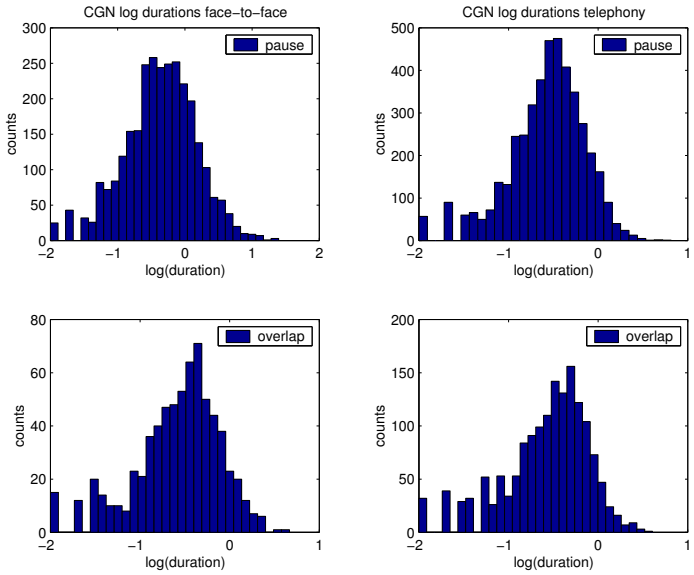


Figure 2. Histograms of the log-durations of face-to-face (left) and telephony (right), and for pauses (top) and overlaps (bottom). The bin size is 0.1. The number of data points for the histograms are 2908, 4375, 1569, and 694, for the panels, clockwise, starting from the left upper panel. For face-to-face data, 27 data points are zero and therefore not plotted. For telephone conversations, this number is 49.

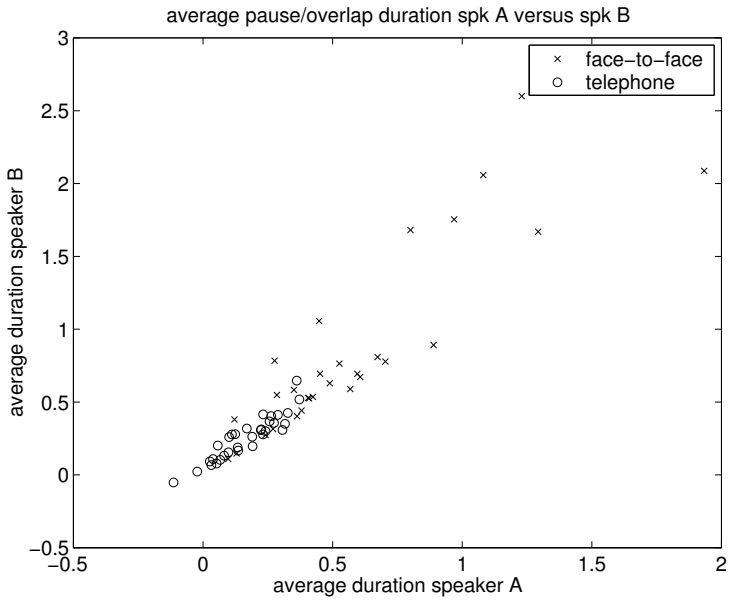


Figure 3. Scatter plot of the average pause duration for speaker A versus speaker B, for both telephone and face-to-face dialogues. Each point represents a dialogue session.

Back-channels or *continuers* are brief responses signaling the interlocutor is still "with the speaker". A continuer can be seen as a signal that the speaker *passes up the opportunity* for taking the next turn (Schegloff, 1982). The function of 'turn' B with reference to turn A is to ACKNOWLEDGE (by means of a responsive word or phrase, a speaker sound), or to DIRECT/REDIRECT (ask to continue (e.g. *dus [so]* ..)). In general, back-channels lack propositional content. If an utterance does have propositional content, it always refers back to something discussed earlier. *Cooperative* turn-taking may take place with a mutual understanding that the turn temporarily shifts to speaker B, only to be handed back to A. Such turn shifts are typically induced by speaker A asking for information or whether B agrees. Turn claims are generally longer stretches of speech; shorter stretches usually concern turn claims that are unsuccessful and abandoned.

## 3 Discussion and conclusion

We have defined the turn concept on the basis of annotations on an utterance-by-utterance basis, in conjunction with data on the start and ending times of the utterances. We realize that a more functional, in-depth account of the turn taking mechanism must be based on an analysis of the material on discourse level, in which the utterances are annotated with respect to their communicative function in context, much like the preliminary analysis in the previous section. Studies suggest a major role for syntax and of prosodic factors for turn-keeping (e.g. Koiso et al., 1998). However, syntactic analysis of spontaneous speech is far from being completely un-

derstood, while detailed prosodic annotation of spontaneous conversations is presently not feasible, due to the time-consuming nature of such an enterprise.

We suggest two plausible explanations for the shorter between-turn pauses in telephone conversations. First, interlocutors in face-to-face interaction have many ways to convey to their partner that they are still involved in the interaction; e.g. by displaying a 'thinking" facial expression. In a telephone conversation speakers must resort to audible signals to indicate that they are still involved in the interaction. Second, in telephone dialogues the conversation is usually the only task the interlocutors are involved in. In face-to-face interaction they can also be engaged in additional tasks, which can by itself provide an account for the longer delay between turns.

**Acknowledgements**

The work of Louis ten Bosch and Jan Peter de Ruiter is made possible by the European IST project COMIC (IST-2001-32311).

## References

Burger, S. (1997). *Transliteration Spontansprachlicher Daten, Lexikon der Transliterationskonventionen in Verbmobil II*. Munich, Verbmobil Technical Report 56-97.

Caspers, J. (2001). Testing the perceptual relevance of syntactic completion and melodic configuration for turn-taking in Dutch. *Proceedings Eurospeech Conference,* pp. 1395-1398.

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

Koiso, H., Horiuchi, Y., Tutiya, S., Ichikawa, A., and Den, Y. (1998). An analysis of turn taking and backchannels based on prosodic and syntactic features in japanese Map Task dialogs. *Language and Speech* 41(3-4), pp. 295-321.

Oostdijk, N., et al. (2002). Het Corpus Gesproken Nederlands. Collection of papers about the Corpus Gesproken Nederlands. LOT Summer School, Netherlands Graduate School of Linguistics, 2002.

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

Schegloff, E.A. (1982). Discourse as an interactional achievement: Some uses of 'uh huh' and other things that come between sentences. In D. Tannen, editor, *Analyzing Discourse: Text and Talk*, pages 71-93. Georgetown University Press, Washington, D.C.

Selting, M. (1996). On the interplay of syntax and prosody in the constitution of turn. Constructional units and turns in conversation, *Pragmatics* 6, pp. 357-388.

Weilhammer, K., and Rabold, S. (2003). Durational aspects in Turn Taking. *Proceedings of the International Conference of Phonetic Sciences*, Barcelona, Spain.