Available online at www.sciencedirect.com
SCIENCE DIRECT®
Speech Communication 47 (2005) 80–86
SPEECH COMMUNICATION
www.elsevier.com/locate/specom

On temporal aspects of turn taking in conversational dialogues

Louis ten Bosch *, Nelleke Oostdijk, Lou Boves

Department of Linguistics, Radboud University Nijmegen, P.O. Box 9103, 6500 HD Nijmegen, The Netherlands

Received 7 April 2005; received in revised form 7 April 2005; accepted 23 May 2005

Abstract

In this short communication we show how shallow annotations in large speech corpora can be used to derive data about the temporal aspects of turn taking. Within the limitations of such a speech corpus, we show that the average durations of between-turn pauses made by speakers in a dyad are statistically related, and our data suggest the existence of gender effects in the temporal aspects of turn taking. Also, clear differences in turn taking behaviour between face-to-face and telephone dialogues can be detected using shallow analyses.

We discuss the most important limitations imposed by the shallowness of the annotations in large corpora, and the possibility for enriching those annotations in a semi-automatic iterative manner.

Keywords: Temporal structure; Spontaneous speech; Dialogues; Turn taking phenomena

1. Introduction

Conversational dialogues show a wide range of turn taking phenomena. In addition to 'canonical' turn switches, where speaker B takes over after speaker A relinquished the floor, overlaps, back-channels (Schegloff, 1982) and interrupts are common events. Moreover, utterances that are syntactically incomplete or ill-formed under any generally accepted grammar abound in spontaneous conversations. This may make it challenging to establish whether a speaker is really prepared to relinquish the floor when he or she produces, e.g., a silent pause. During the last decades, scientists from a range of disciplines have tried to come to grips with these phenomena, with special focus on the role of syntactic, prosodic and semantic/pragmatic factors (Allport et al., 1972; Sacks et al., 1974; Schegloff, 1982; Clark, 1996; Ford and Thompson, 1996; Ward, 1999; Ward and Tsukahara, 2000; Caspers, 2001, 2003; Selting, 1996, 2000). Most of the older studies are based on small samples of rather formal dialogues, which

* Corresponding author. Tel.: + 31 24 3616069; fax: + 31 24 3612907.
E-mail addresses: l.tenbosch@let.ru.nl (L. ten Bosch), n.oostdijk@let.ru.nl (N. Oostdijk), boves@let.ru.nl (L. Boves).

0167-6393/$ - see front matter © 2005 Elsevier B.V. All rights reserved.
doi:10.1016/j.specom.2005.05.009

tends to make their theoretical appeal greater than their empirical and descriptive basis.

The recent advent of large corpora with conversational dialogues has opened up opportunities for novel empirical research into turn taking phenomena, e.g., Switchboard (Godfrey et al., 1992) and the Spoken Dutch Corpus (CGN) (Oostdijk, 2002). However, the size of these corpora comes with a cost: the level of annotation detail is necessarily quite limited. For example, in CGN no information about turns, or information concerning the function of utterances in a context of the dialogue is available. In parts of the Verbmobil corpus (Wahlster, 1997) turn-like units could be identified thanks to the constraints imposed on the conversations (Weilhammer et al., 2000). In general, however, few large corpora come with functional information about turns and utterances. Therefore, without additional annotation these corpora cannot be used directly to test the hypotheses and claims in the literature on turn taking and related phenomena.

Yet, this does not imply that research in this direction is impossible. In this communication we want to show how large corpora with only basic annotations can be used to study temporal phenomena in conversational dialogues. More in particular, we will indicate a statistical dependency in average duration of between-turn pauses between speakers in a telephone dialogue, and differences between face-to-face dialogues and telephone conversations.

We intend to demonstrate that empirical studies on large corpora with only basic annotation can lay the groundwork for further analyses of turn taking in conversational dialogues. Moreover, we suggest that empirical studies can be instrumental in arriving at a general agreement about how turns can be identified and what set of function labels is minimally required in order to permit a robust level of (semi-)automatic analysis, while at the same time maintaining a reasonable degree of reliability.

2. Pauses and overlaps

Several studies have investigated pause and overlap phenomena in conversations (e.g., Jefferson, 1989; Zeller, 1994; Weilhammer and Rabold, 2003). The frequency and distribution of these phenomena is different for monologues on the one hand, and dialogues on the other (Campione and Veronis, 2002) and are dependent on particular modalities (e.g., Duncan and Fiske, 1977; Cassell et al., 1999). Once the concept of 'turn' is defined, one can distinguish silent pauses within turns and between turns. Within a turn pauses can occur either within an utterance, or between two utterances that are part of the same turn. However, turns from different speakers need not be separated by a silent pause. It is quite usual for turns to overlap to some extent (e.g. Sellen, 1995), and this can happen without conveying the impression that speakers are being impolite, interrupting each other and openly competing for keeping or getting the floor. On the contrary, most persons will fall silent in the absence of supportive backchannel noises from the interlocutors, which may overlap completely with their speech, or just fill their silent pauses (see e.g., Schegloff, 1982; Caspers, 2001).

Filled pauses, i.e., stretches of speech that are transcribed as 'ehm', 'eh', 'uhm', etc., where the exact transcription may depend on the language, are more complex. In this communication, we simplify the pause/overlap analysis by considering filled pauses as words on a par with lexical words, for the simple reason that if a speaker produces a filled pause, this is most likely a sign that the current turn is not yet over. Both 'ums' and 'uhs' indicate the expectation of upcoming delays (albeit in a slightly different way, Clark and Fox Tree, 2002). Furthermore, if the filled pause is the first 'word' in a turn, this can be considered to be the start of the turn. Thus, in the analysis of temporal phenomena in turn taking presented in this communication we only take silent pauses and overlaps into account. This is in line with choices in earlier studies (such as Campione and Veronis, 2002; Weilhammer and Rabold, 2003).

Disregarding backchannels for the time being, overlapping speech can represent at least two functionally different situations (see Roger et al., 1988, for a full account). First, the verbal content, syntactic structure or the prosody of the ongoing utterance might indicate a transition-relevance

place (e.g., Selting, 2000) or an explicit intention to relinquish the floor. In such a case, turn taking can be considered an unmarked event. Second, the ongoing utterance may use the same set of devices (vocabulary, syntax and prosody) to convey the intention to *keep* the floor. In this case, taking over the floor will most likely be perceived of as an interruption. Although turn changes with 'overlaps' can be functionally different, in terms of the temporal structure exploited here they are considered the same: there is a time interval in which both speakers are talking.

3. Labelling method and data

This communication is based on analyses of part of the spontaneous conversations in the Spoken Dutch Corpus. All verbatim transcriptions in the corpus contain a basic level of punctuation that allows separating the sequence of words into utterances, without any correspondence to a 'grammatical' structure. The annotation does not group utterances into higher level discourse units, such as turns in a conversation. For this reason we have defined 'turns' as stretches of one or more utterances that are not interrupted by another speaker. This operational definition of the concept of turn makes it possible to distinguish between turn internal pauses on the one hand and pauses at the boundary of a turn on the other hand.

For this study, we have taken the spontaneous dialogues (telephone dialogues, and face-to-face dialogues) in the corpus that come with manually verified word segmentation. In this part of the corpus it is possible to automatically derive data on within-turn and between-turn pauses, as well as the duration of overlaps between adjacent turns. The analysis is based on a proposal made by Weilhammer and Rabold (2003); it is summarized in Fig. 1 in a simplified form that leaves out subtypes. All turn changes between the speakers in a dialogue have been classified using this labelling scheme.

4. Results

We first present global data obtained from an analysis of 93 telephone dialogues (about 15.1 h of speech). In Table 1 we have tabulated means, medians and standard deviations of three types of silent pauses: within utterances, between utterances in one turn, and at turn transition points. Pauses between utterances in one turn appear significantly ($p < 0.05$) longer than utterance internal pauses and pauses at turn changes, in agreement with earlier findings (Jefferson, 1989; Campione and Veronis, 2002).

4.1. Accommodation of pause durations

Next, we examined the average duration of between-turn pauses (related to 'speaker change' in Fig. 1) for both speakers in all 93 telephone dialogues. To that end, the inter-turn pauses were attributed to the speaker following that pause. In Fig. 2, each of the 93 dialogues is represented by a point, of which the coordinates are determined by the average duration of the inter-turn pauses of the speakers in that dyad. The correlation as shown in this data set is 0.47 (significantly different from 0: Fisher's $z > 3$, $p < 0.001$). This means that the observed patterning is significantly different

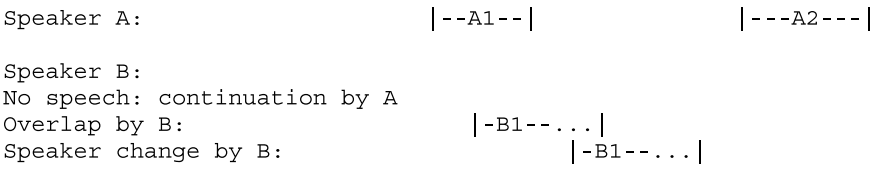

Fig. 1. Simplified overview of temporal patterning of utterances. Time flow is represented from left to right; the scheme refers to each moment when an utterance (here labelled A1) is completed by speaker A. The symbols A2 and B1 denote the next utterance by speaker A and by speaker B, respectively. At the moment A has finished utterance A1, three different situations are distinguished, here labelled as 'continuation', 'overlap', or 'pause', depending on the temporal organisation of B1 and A2 relative to A1. Sub-cases are not shown.

Table 1
Summary information about the distributions of three types of pauses, measured over all 93 telephone dialogues

Pause	Number	Mean	Median	Std. dev.
Within utterance	810	0.30	0.28	0.21
Continuations	1860	0.52	0.45	0.38
Speaker changes	6790	0.38	0.33	0.31

Means, medians and standard deviations (std. dev.) are in seconds.

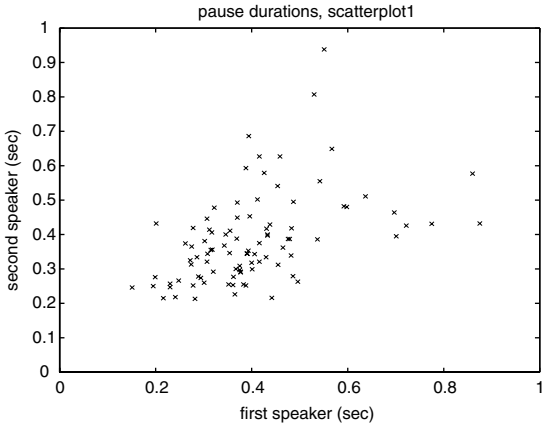

Fig. 2. Scatter diagram of average pause durations for 93 dialogues. The axes represent average durations in seconds. Each point represents one dialogue.

from a patterning that would be obtained by random speaker-to-speaker assignments into dyads. It can be shown that this difference is not an artefact of the assignment of speakers *within* dyads, but that it remains when speakers within dyads are interchanged. Thus, it suggests a form of accommodation with respect to inter-turn pause duration, which can be interpreted in the context of more general results on accommodation in dialogues and conversations (Tannen, 1989; Giles et al., 1992; Garrod and Pickering, 2004).

For durations of pauses between utterances *within* turns, the correlation between average pause durations is much smaller ($r < 0.15$, $z = 1.43$, $p < 0.076$, i.e., not significantly different from 0). That suggests that the trend shown in Fig. 2 cannot be generalized to pauses between utterances within turns. In other words, it seems that between-speaker accommodation of the temporal structure in the form as evaluated here is mainly limited to turn taking phenomena, i.e., the timing aspect that is most directly related to the collaboration between the speakers. However, accommodation is certainly not the only phenomenon that may give rise to a medium-sized correlation between the duration of inter-turn pauses. Another potential cause that comes to mind is the overall character of a dialogue. It may well be that a lively conversation on some pleasant shared experience gives rise to shorter pauses than a discussion on some difficult problem. Adaptation of pause duration in the context of human–machine interaction is described in Oviatt et al. (2004), who found young children adapting their pause durations to the introvert or extravert character of an avatar in a human–machine dialogue.

In addition, our data are definitely influenced by the definition of 'silent pause'. There is no generally agreed definition of the acoustic properties of the signal that must lead to the detection of a pause. In their analysis of pause durations, Campione and Veronis (2002) explicitly mention the potentially drastic effects of the use of duration threshold values, which are commonly applied for practical reasons, but may easily lead to biased analysis results. The current analysis strictly takes the word segmentation as starting point, leaving the ultimate interpretation of 'short pause' to the labelling procedure. In the CGN data, the proportion of very short pauses is substantial, as is shown in Table 2. About 14% of all the pauses (i.e., pauses within utterance, pauses between utterance, and pauses between turns) are shorter than 150 ms, a lower bound that often plays a role in

Table 2
Cumulative distribution of all short pauses (i.e., within utterance, between utterances, and between turns) annotated in dialogues per telephone in the sub-corpus under analysis

Threshold (s)	(Cumul.) proportion (%)
<0.05	2.5
<0.10	4.0
<0.15	13.5
<0.35	49.0
<0.50	60.0

pause duration statistics (see also Campione and Veronis, 2002). However, given the manual segmentation of the speech material, it is guaranteed that none of the short silent intervals mistake stop closures for pauses. At the same time it is true that a 150 ms pause preceded by a certain prosodic contour may be much more salient perceptually than a longer pause in the absence of prosodic boundary cues.

4.2. Overlap statistics

Overlaps form an inherently more difficult category than pauses, both from a functional point of view (Roger et al., 1988) and from a labelling perspective: the definition of an overlap as the event 'speaker B already speaks before speaker A stops speaking' is not always unambiguous. In the following example, the question is how the situation after B1 finishes must be labelled. Is A2 overlapping B1? Weilhammer and Rabold (2003) classify this as an overlap between B1 and A2, but the phenomenon cannot be unambiguously attributed to any of the speakers.

```
Speaker A:  |--A1--|          |----A2--------|
Speaker B:                    |----B1---|
```

To avoid these interpretation issues, we studied the proportion of overlaps compared to the total number of turn changes, rather than the statistics of overlap durations. Using two different statistical tests, we observed an influence of gender on this proportion: the 10 male–male telephone conversations show a higher proportion of overlaps, compared to the 36 female–female conversations (the difference is significant at the level $p = 0.05$: $t = 1.69$, $\mathrm{df} = 44$, one-tailed; Wilcoxon $Q = 119$, $n_A = 10$, $n_B = 36$). The male–female conversations were not taken into account in this comparison.

4.3. Comparison between telephone and face-to-face conversations

The face-to-face conversations in CGN differ in many respects from the telephone conversations. For example, while most telephone conversations were friendly chats, a substantial proportion of the face-to-face conversations is about games that are being played, or other tasks that are being performed during the conversations (such as laying tables). Moreover, several studies have indicated substantial changes in interaction when more than one modality is available for communication (e.g., Duncan and Fiske, 1977; Cassell et al., 1999). By comparing 29 face-to-face dialogues and 32 telephone dialogues (a subset of the 93 dialogues used in the previous section; in total, these sets contain 301 and 306 min, respectively), it appears that face-to-face dialogues have a significantly lower proportion of overlaps. Table 3 shows the difference of number of overlaps and pauses at turn changes for telephone and face-to-face dialogues. The proportion of overlaps decreases from 52% for telephone to 44% in the face-to-face setting (significant, $\chi^2 > 57$, $\mathrm{df} = 1$, $p < 0.001$). Evidently, the absolute number of overlaps and pauses per time unit is substantially lower in the face-to-face conversations than in telephone dialogues.

5. Discussion and conclusion

The advent of large corpora with conversational dialogues has opened up opportunities for empirical research into turn taking phenomena. In such corpora, the level of annotation detail is often quite limited, mainly because of the costs that are associated with a detailed labelling. In most cases, these corpora only contain shallow information about turns or about the function of

Table 3
Counts and percentages of speaker changes with pauses and overlaps, for telephone and face-to-face dialogues

Case	Telephone (counts)	Telephone (%)	Face-to-face (counts)	Face-to-face (%)
Overlap	2697	52	1589	44
Pauses at speaker change	2491	48	2040	56
Total	5188	100	3629	100

utterances in the context of the dialogue. In this paper, we indicated how large corpora with only a basic level of annotation can be used to study temporal phenomena in conversational dialogues. The analysis of pauses and overlaps has been simplified by considering filled pauses as lexical words (Clark and Fox Tree, 2002), and by focussing on silent pauses and overlaps.

We have shown that there is a significant correlation between the average durations of between-turn pauses (related to 'speaker change' in Fig. 1) in telephone dialogues produced by the two interlocutors. This suggests a form of accommodation with respect to inter-turn pause duration, which can be interpreted in the context of more general results on accommodation in dialogues and conversations (Tannen, 1989; Garrod and Pickering, 2004). Another possible explanation is that average durations of inter-turn pauses mainly depend on the overall character of a dialogue (e.g., vivid with many interactions, or a lower-pace discussion about a difficult problem). Clear differences were found between face-to-face and telephone conversations, supporting findings in earlier studies (e.g., Cassell et al., 1999).

From our work it also becomes clear that it is difficult to relate the 'acoustic' and linguistic data that one can extract (semi-)automatically from shallow annotations to more formal theories about turn taking. Even if the verbatim transcription is carried out according to a protocol that is pragmatically defined and linguistically adequate (see Oostdijk, 2002), there is no straightforward way to interpret these annotations in terms of the function of utterances and turns in the context of theories of discourse and conversational analysis. For example, the application of a turn labelling scheme such as proposed by Weilhammer and Rabold (2003) is not obvious if one aims to distinguish a 'pure' backchannel from a user utterance with verbal contents, e.g., of words such as 'sure' and 'indeed'. However, from the experience gained with the use of CGN, we are confident that it is possible to develop semi-automatic procedures for adding annotation related to the discourse function of utterances and turns. If words like 'indeed', 'yes', etc., fully overlap with the interlocutor's speech, they are very likely to function as backchannels. The same may be true if there is no temporal overlap, and the preceding turn is not a question or request. The classification of utterances as questions/requests can be accomplished on the basis of a combination of the verbal contents, punctuation marks and the prosody. Finally, Machine Learning techniques should be able to bootstrap such labelling procedures from reasonably small amounts of training data (Day et al., 1997; Fernandez et al., 2005).

Acknowledgements

The contribution by Louis ten Bosch has been made possible by the European IST project COMIC (IST-2001-32311). Thanks are due to Peter Beinema for preparing the data that were used in the analysis of pause durations and distributions of pauses and overlaps in the 93-dialogue corpus.

References

Allport, D.A., Antonis, B., Reynolds, P., 1972. On the division of attention: a disproof of the single channel hypothesis. Q. J. Exp. Psych. 24, 225–235.

Campione, E., Veronis, J., 2002. A large-scale multilingual study of silent pause duration. ESCA-workshop on speech prosody, April 2002, Aix-en-Provence, pp. 199–202.

Caspers, J., 2001. Testing the perceptual relevance of syntactic completion and melodic configuration for turn-taking in Dutch. Proc. Eurospeech, pp. 1395–1398.

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

Cassell, J., Torres, O., Prevost, S., 1999. Turn taking vs. discourse structure: how best to model multimodal conversation. In: Wilks, Y. (Ed.), Machine Conversations. Kluwer.

Clark, H., 1996. Using Language. Cambridge University Press.

Clark, H.H., Fox Tree, J.E., 2002. Using uh and um in spontaneous speech. Cognition 84, 73–111.

Day, D., Aberdeen, J., Hirschman, L., Kozierok, R., Robinson, P., Vilain, M., 1997. Mixed-initiative development of language processing systems. Proc. 5th Conf. on Appl. Nat. Language Process, pp. 348–355.

Duncan, S.D., Fiske, D.W., 1977. Face-to-Face Interaction: Research, Methods and Theory. Lawrence Erlbaum, Hillsdale, NJ.

Fernandez, R., Ginzburg, J., Lappin, S., 2005. Automatic bare sluice disambiguation in dialogue. Proc. of IWCS-6, pp. 115–127.

Ford, C.E., Thompson, S.A., 1996. Interactional units in conversation: syntactic, intonational, and pragmatic resources for the management of turns. In: Ochs, E.,

Schegloff, Thompson, S.A. (Eds.), 2003, Interaction and Grammar. Cambridge University Press, Cambridge, pp. 134–184.

Garrod, S., Pickering, M.J., 2004. Why is conversation so easy? Trends Cognitive Sci. 8 (1), 8–11.

Giles, H., Coupland, N., Coupland, J., 1992. Accommodation theory. In: Giles et al. (Eds.), Contexts of Accommodation. Communication, Context and Consequences. Cambridge University Press, pp. 1–68.

Godfrey, J.J., Holliman, E.C., McDaniel, J., 1992. SWITCHBOARD: Telephone speech corpus for research and development. In: Proc. IEEE Internat. Conf. on Acoustics, Speech and Signal Process. (ICASSP), San Francisco, USA, pp. 517–520.

Jefferson, G., 1989. Preliminary notes on a possible metric which provides for a 'standard maximum' silence of approximately one second in conversation. In: Roger, D., Bull, P. (Eds.), Conversation and Interdisciplinary Perspective, vol. 3. Multilingual Matters Ltd., Clevedon, pp. 166–196.

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

Oviatt, S., Darves, C., Coulston, R., 2004. Toward adaptive conversational interfaces: modeling speech convergence with animated personas. ACM Trans. Computer–Human Interaction (TOCHI) 11 (3).

Roger, D., Bull, P., Smith, S., 1988. The development of a comprehensive system for classifying interruptions. J. Lang. Social Psychol. 7, 27–34.

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

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

Sellen, A.J., 1995. Remote conversations: the effect of mediating talk with technology. Hum. Comput. Interact. 10, 401–444.

Selting, M., 1996. On the interplay of syntax and prosody in the constitution of turn. Constructional units and turns in conversation. Pragmatics 6, 357–388.

Selting, M., 2000. The construction of units in conversational talk. Lang. Soc. 29, 477–517.

Tannen, D., 1989. Talking Voices: Repetition, Dialogue and Imagery in Conversational Discourse. Cambridge University Press.

Wahlster, W., 1997. Verbmobil: Uebersetzung von Verhandlungsdialogen. Verbmobil report 01–93. Available from: <http://www.defki.de/cgi-bin7verbmobil>.

Ward, N., 1999. Low-pitch regions as dialog signals? Evidence from dialog-act and lexical correlates in natural conversation. In: Proc. ESCA Workshop on Dialogue and Prosody, Veldhoven, the Netherlands, pp. 83–88.

Ward, N., Tsukahara, W., 2000. Prosodic features which cue back-channel responses in English and Japanese. J. Pragmatics 23, 1177–1207.

Weilhammer, K., Oppermann, D., Burger, S., 2000. The influence of scenario constraints on the spontaneity of speech. A comparison of dialogue corpora. In: Proc. LREC (cd-rom).

Weilhammer, K., Rabold, S., 2003. Durational aspects in Turn Taking. In: Proc. Internat. Conf. of Phonetic Sci., Barcelona, Spain (cd-rom).

Zeller, B., 1994. Pauses and the temporal structure of speech. In: Keller, E. (Ed.), Fundamentals of Speech Synthesis and Speech Recognition. John Wiley, Chichester, pp. 41–62.