

Timelines, transcription and turn-taking: a corpus-driven approach to simultaneous speech

Richard Forsyth¹, David Clarke¹, Phoenix Lam²

¹ School of Psychology, University of Nottingham, U.K.

² Department of English, Hong Kong Polytechnic University, Hong Kong.

rsf@psychology.nottingham.ac.uk

ddc@psychology.nottingham.ac.uk

eg.phoenix@polyu.edu.hk

Abstract

Representations of spoken discourse must accommodate the phenomenon of simultaneous speech. Linguists and other social scientists have employed numerous transcription conventions for exhibiting the temporal interleaving of multi-speaker talk (e.g. Atkinson & Heritage, 1984; Schifffrin, 1994; Leech *et al.*, 1995; Carter, 2004; MICASE, 2004). Most of these conventions are mutually incompatible. The existence of many different systems is evidence that representing turn-taking in natural dialogue remains a problematic issue.

The present study discusses a novel orthographic transcription layout which records how participants contribute to the stream of spoken events based on word timings. To test this method, the Maptask corpus (Anderson *et al.*, 1991) was used because it contains unusually precise information on the timings of vocal events. Using the term *vocable* to denote words and a small number of short phrases, every vocable in the Maptask corpus has its onset and ending time recorded. This painstaking attention to timing permits examination of overlapping speech in greater detail than has been customary.

A non-standard talk-division format was generated by a simple algorithm that sorts each vocable in temporal order and appends it to the current line if it was produced by the same speaker as the last one, otherwise prints it on a new line preceded by a time-stamp and speaker label. Thus the alternation of speakers is not imposed by a transcriber's intuition about what constitutes a turn but emerges from the empirical data. This offers an *etic* perspective on turn-taking, in contrast to the *emic* perspective of traditional approaches. It tends to highlight the prevalence of "echoing" in the joint production of dialogue. Moreover, lengths of speech segments and inter-speaker intervals as defined by this procedure showed highly significant associations with a number of contextual and interactional variables, indicating that this approach can yield analytic as well as representational benefits.

1. Introduction

Representations of spoken discourse must allow for the phenomenon of simultaneous speech. Linguists and other social scientists have employed numerous transcription conventions for exhibiting the temporal interleaving of multi-speaker talk (e.g. Boden

& Zimmermann, 1991; Schiffrin, 1994; Leech *et al.*, 1995; Carter, 2004; MICASE, 2004). Most of these conventions are mutually incompatible. Researchers have used a variety of symbols to mark the start and end of simultaneous speech, including braces, brackets, hash signs, and XML-style tags; in addition, they may or may not attempt to align overlapping speech segments vertically on the page. In other words, different researchers make different compromises with the ideal of a sequence of complete non-overlapping utterances.

The great majority of transcription schemes derive ultimately from "the familiar conventions of the playscript" (Payne, 1995: 206). The fact that different researchers make different compromises with this ideal is evidence that representation of turn-taking in natural dialogue remains a problematic issue.

For the English language, two broad families of transcription conventions can be discerned, one deriving from the field of Conversation Analysis (Sachs *et al.*, 1974) and the other arising from the needs of Corpus Linguistics (Sinclair, 1991). However, even within these traditions there are variations. For example, Schiffrin (1994), writing within the Conversation Analysis tradition, gives three different styles of transcription in an appendix, while the collection edited by Leech *et al.* (1995), written from a Corpus Linguistics perspective, contains at least four systems of transcription for speech.

The present study was performed to find out what might be gained by jettisoning the idea of transcript as playscript, and representing in a straightforward manner the order in which participants contribute to the stream of spoken events. For this purpose, the Maptask corpus (Anderson *et al.*, 1991) was used because it contains unusually precise information on the timings of vocal events. Using the term *vocable* to denote words and certain short phrases spoken without internal silence (such as "do you", "going to" and "you know"), every single vocable in the Maptask corpus (a total of over 150,000) has its onset and ending time recorded, accurate to the nearest 1/100th of a second. This painstaking attention to timing permits the examination of the interleaving and overlapping of spoken contributions in greater detail than has previously been customary.

2. The Maptask corpus

The Edinburgh Maptask corpus (Anderson *et al.*, 1991) consists of a total of 128 dyadic dialogues. These were produced by 64 undergraduates from the University of Glasgow, 32 females and 32 males. Each dyad was given a task, the "Map Task" (Brown *et al.*, 1984). In this task both participants have a sketch-map with several landmarks on it. They cannot see each other's maps. One participant, the *instruction giver*, has a path printed on his or her map, and is required to guide the other participant, the *instruction follower*, in drawing that path on the other map, which has no path marked on it. These two maps are similar but not identical, and the participants are told this fact. The number of (paired) maps used was 16 and in each pair the number of mismatching features between giver's and follower's map was equivalent. Each participant took part in four conversations, twice as an instruction giver (with both a familiar and an unfamiliar partner) and twice as an instruction follower (with both a familiar and an unfamiliar partner). In half of the conversations

the speakers could make eye contact while in the other half a cardboard screen prevented them from seeing each other.

A noteworthy feature of this task is that it provides a quantifiable measure of communicative success. This is realized as a deviation score, which is computed as the area, in square centimetres, between the route as shown on the instruction giver's map and the path drawn on the instruction follower's map. Thus higher values indicate lesser success in the task. Having an outcome measure, such as this one, made it possible for the present study to examine associations between patterns of spoken interchanges and communicative success.

More generally, the present study focuses on examining the relationships between certain features of spoken contributions, particularly their lengths, and the outcome variable (path deviation score) as well as a number of other contextual variables describing attributes of the speakers and/or interaction, as listed in Table 1, below.

Name	Description	Range
Ages	Age of participants in years.	17 .. 30
Eye-contact	Whether participants could see each other's faces	0, 1
Familiarity	Whether speakers were already acquainted	0, 1
Gender	Sexes of participants	F, M
Role	Role in the interaction, giver or follower	F, G
Outcome	Path deviation score	4 .. 227

Table 1. Principal Maptask Variables.

Little attention was given to the ages of the participants, since these fell in a rather narrow range: more than 89% were aged 18-22.

3. A talk-slicing technique

A non-standard talk-division format was generated by an algorithm that simply sorted each vocable into temporal order, according to its onset time, and processed all the vocables sequentially, printing them out according to the following rule:

- if the current vocable was produced by the same speaker as the last one, append it to the current line;
- else print the current vocable on a new line preceded by a time stamp and speaker label.

Thus the arrangement of contributions by speaker is not imposed by preconceptions about well-behaved spoken interaction but emerges from the empirical data.

Provided that each vocable is associated with an accurate onset time, this technique is extremely simple to implement. (Admittedly, accurate timings at the word level are still rare but they are becoming less so as automatic word-boundary detection software improves.) For the present investigation this procedure was implemented by a program in the Python language. This is referred to as TST1 (Talk Slicing Technique, number 1) in what follows.

The TST1 program also offers the option of inserting another column between the time stamp and the speaker label, which gives the time interval in seconds from the ending of the last vocable of the previous speaker to the start of the first vocable of the current speaker. If this number is negative, it indicates an overlap of voices (see, for example, Table 3).

4. Talk slicing on the page

4.1 Simple talk slicing

The visual effect of this process can be illustrated by the extract below, from Maptask dialogue q4nc2. Table 2 gives the dialogue first as segmented by a human observer, followed by the segmentation produced by the TST1 system. Two noticeable features are that the TST1 layout tends to include more stretches of "vertical" talk, and that it highlights the phenomenon of *echoing*, where words and short phrases are passed from one speaker to another.

Segmentation by human transcriber	
G:	okay , we're s-- ... we're starting in the top left-hand corner .
F:	uh-huh , uh-huh ... at the start .
G:	ehm a-- ... above the diamond mine .
F:	uh-huh that's right .
G:	ehm come straight down .
F:	straight down ?
G:	straight down .
F:	straight down yeah ?
G:	straight down .
F:	uh-huh , how far , out , oh well i'll move this over then ... okay , okay then , right ... uh-huh .
G:	is i-- ?
F:	okay , right okay .
G:	do you have a , can i ... can you , do you do you have a field station marked on the left-hand side ?
F:	uh-huh , do i go down that far ?
G:	go no , go down halfway between the ... the .
F:	the diamond mine ... and the field station .
G:	the diamond mine and the field station .
F:	okay .
G:	and then turn to ... the right .
F:	the right ... okay , how far ?
G:	i'm g-- ... ca-- is there ... a mountain in the top ?
F:	the highest viewpoint ?
G:	the highest viewpoint .
F:	uh-huh .
G:	are springboks marked on the map ?
F:	ehm ... no .
G:	no , well come ... over to towards the ... the highest viewpoint ... come s-- come , make sure you've come halfway between ... the diamond mine and the field station .
Segmentation produced by TST1 procedure	
SEGMENTATION BY TST1	
1.48	G: okay we're
2.15	F: uh-huh
2.27	G: s-- we're starting in the top left-hand corner
4.58	F: uh-huh at
5.22	G: ehm
5.34	F: the start
5.47	G: a-- above the diamond mine
7.02	F: uh-huh that's right
8.29	G: ehm come straight down
10.67	F: straight down
11.16	G: straight down
12.76	F: straight down yeah

13.62 G: straight down
 14.33 F: uh-huh how far out oh well i'll move this over then okay okay
 then right uh-huh
 28.29 G: is i--
 28.60 F: okay right okay
 30.45 G: do you have a can i can you do you do you have a field station
 marked on
 35.45 F: uh-huh
 35.53 G: the left-hand side
 36.34 F: do i go down that far
 37.18 G: go no go down halfway between the the
 41.94 F: the diamond
 42.27 G: the diamond
 42.57 F: mine
 42.73 G: mine and
 43.08 F: and
 43.21 G: the
 43.24 F: the
 43.27 G: field
 43.30 F: field
 43.53 G: station
 43.54 F: station okay
 44.87 G: and then turn to the right
 48.61 F: the right okay how far
 50.90 G: i'm g-- ca-- is there a mountain in the top
 54.66 F: the highest viewpoint
 55.18 G: the highest viewpoint
 55.86 F: uh-huh
 56.23 G: are springboks marked on the map
 58.38 F: ehm no
 59.55 G: no well come over to towards the the highest viewpoint come s--
 come make sure you've come halfway between the diamond mine and the field
 station

Table 2. The first 60 seconds of Maptask dialogue q4nc2.

In this extract TST1 divides the conversation into more contributions (41 versus 27) than the turn-division of the expert human transcriber. Therefore the contributions are shorter, on average. This is typical.

Impressionistically, it can be said that the TST1 format emphasizes the joint production of talk rather more than is conventional and foregrounds the frequent short utterances of one or two words that are often omitted in non-specialist transcription (e.g. journalistic reporting) and would mostly be treated as backchannel contributions by linguists. If such vocal signals are noted in conventional transcriptions, they tend to be tacked onto the nearest "proper" (content-bearing) utterance, so their temporal placement is often imprecise.

Of course the two versions are not dramatically different, nor would we wish them to be, since simultaneous speech, though frequent in conversation, is not the norm. Thus there are plenty of examples where the two methods divide the stream of speech at the same points. Nevertheless, it can be said that, relatively speaking, TST1 gives prominence to short contributions of one, two or three vocables. These are, in a sense, given equal status with the longer, more linguistically salient utterances.

Whether this "upgrading" of short contributions which may be paralinguistic affirmations, interjections or aborted attempts at content-bearing utterances makes the dynamics of the interaction stand out more clearly on the page is for the human reader to judge. Whether it offers an analytical payoff is the subject of section 5.

4.2 Talk slicing showing overlaps

As an option, the TST1 program can insert another column of figures between the onset time and the speaker code. This gives the time interval in seconds from the ending of the last word of the previous speaker to the start of the first word of the current speaker's contribution. If this interval is negative, it indicates an overlap. An example is shown in Table 3.

Conversation divided by human observer with overlapping speech in light blue.		
<u>GIVER:</u>		do you have a , can i ... can you , do you do you have a field station marked on the left-hand side ?
<u>FOLLOWER:</u>		uh-huh , do i go down that far ?
<u>GIVER:</u>		go no , go down ... halfway between the ... the .
<u>FOLLOWER:</u>		the diamond mine ... and the field station .
<u>GIVER:</u>		the diamond mine and the field station .
<u>FOLLOWER:</u>		okay .
<u>GIVER:</u>		and then turn ... to ... the right .
<u>FOLLOWER:</u>		the right ... okay , how far ?
Conversation divided by TST1 with timestamps in column 1 and inter-speaker interval in column 2. Overlaps are coloured red; gaps longer than 1 second are coloured blue; "normal" intervals are coloured green.		
30.45	0.16	G: do you have a can i can you do you do you have a field station marked on
35.45	-0.08	F: uh-huh
35.53	-0.32	G: the left-hand side
36.34	-0.17	F: do i go down that far
37.18	-0.27	G: go no go down halfway between the the
41.94	1.20	F: the diamond
42.27	-0.30	G: the diamond
42.57	-0.15	F: mine
42.73	-0.32	G: mine and
43.08	-0.12	F: and
43.21	-0.03	G: the
43.24	-0.03	F: the
43.27	-0.03	G: field
43.30	-0.24	F: field
43.53	-0.01	G: station
43.54	-0.52	F: station okay
44.87	0.30	G: and then turn to the right
48.61	0.15	F: the right okay how far

Table 3. A section of dialogue q4nc2 illustrating overlaps.

In this extract, overlapping speech has been highlighted by colour coding. In the conventional transcription overlapping portions of speech are coloured light blue. In the transcription produced by TST1 the second column of numbers shows the inter-slice interval. When this is negative it indicates an overlap, and has been coloured red. Gaps of more than 1 second have been coloured blue.

We argue that this representation gives more information about the temporal interleaving of the participants' speech than conventional layouts, without sacrificing the intelligibility of the linguistic content.

5. Results

In this section we briefly review some numerical characteristics of the dialogues in the Maptask corpus, in order to give background information, and then examine some

of the more striking relationships between length of speakers' contributions and the other variables of interest. To avoid misunderstanding we use the term *contribution* throughout the section to designate a segment of speech as defined by the TST1 procedure, because simpler terms such as "turn" or "utterance" carry a number of linguistic and interpretive connotations which might obscure the nature of our analyses.

5.1 Some basic Statistics of the dialogues in the Maptask corpus

The corpus contains 128 dialogues which between them consist of 26193 contributions containing a total of 153,780 word tokens. Table 4 shows the numerical characteristics of the dialogues in the Maptask corpus in relation to contribution (as defined by TST1), word token and deviation score. Mean and median values are shown, along with standard deviations.

Variable	Mean	Median	Standard deviation
Contributions per dialogue	204.63	169.50	122.70
Word tokens per dialogue	1201.41	1014.00	648.00
Contribution length in word tokens	5.87	3.00	7.37
Contribution length in seconds	1.99	1.01	2.77
Deviation score	71.82	56.00	49.17

Table 4. Basic Statistics of Maptask Variables.

None of these variables is symmetrically distributed, all being skewed to the right.

5.2 Some correlatives of contribution length

In this section we present some major findings on the associations between the four main contextual variables -- eye-contact, familiarity, gender and speaker role -- with the lengths of contributions as defined by TST1.

5.2.1 Instruction followers make shorter contributions than givers. Of the four main contextual features examined the variable with the strongest association with length of speakers' contribution is speaker role. It is obvious from the histograms below (Figure 1) that the contribution lengths of givers and followers differ dramatically. In particular, followers produce far fewer long contributions. Some difference is to be expected since they have different communicative goals; perhaps more worth remarking is that both distributions are **bimodal**. This suggests that each distribution is a composite of two different types of contribution. (The lengths of the contributions in seconds have been subjected to a logarithmic transformation for display purposes in Figure 1, as the original distribution is very strongly skewed to the right.)

The normal curves superimposed show that these timings are definitely not normally distributed: there is a clear preponderance of short contributions, for both parties, though significantly more so for followers ("f:") than givers ("g:").

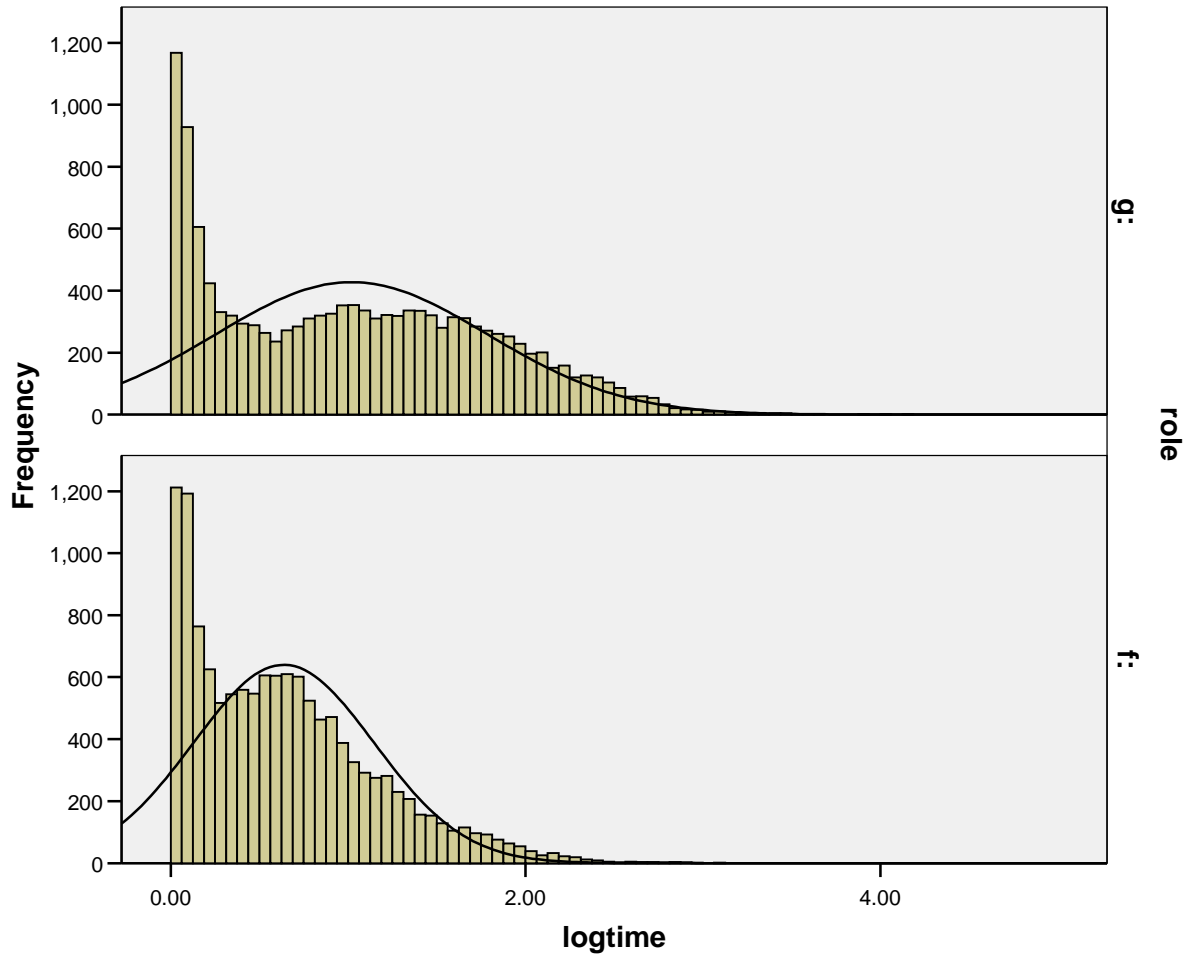


Figure 1. Histograms of $\text{logtime} = \ln(t+1)$ where t is length of contribution in seconds.

As these variables are not normally distributed a non-parametric test, the Wilcoxon rank-sum test¹, was performed to compare contribution lengths between givers and followers. The results for both time in seconds and length in word tokens were very highly significant: for times, Wilcoxon equivalent z -score = -38.35, $p < 0.0005$; for tokens, Wilcoxon equivalent z -score = -49.32, $p < 0.0005$.

Another way of looking at this effect is by tallying the frequency of single-word contributions between the two speaker roles, as presented in Table 5.

Contribution length:	1-Word Contributions	Longer Contributions
Instruction Giver	3767	9359
Instruction Follower	6554	6513

Table 5. Cross-tabulation of Speaker Role and Contribution Length.

Over half (50.2%) of the followers' contributions consisted of just a single word, whereas only 28.7% of the givers' contributions were single words. Thus the odds in favour of a follower's contribution being a 1-word utterance were 2.5 times as great as the odds for a giver's. Statistically, this difference is very highly significant indeed (Chi-squared = 1261.87, $df = 1$, likelihood ratio = 1274.96, $p < 10^{-278}$).

5.2.2 Familiar pairs produce more exchanges than unfamiliar, but they are shorter. There is a clear tendency for pairs who are already acquainted to produce a larger number of contributions per dialogue than those who are unfamiliar. The median number of contributions for the familiar pairs was 191 while for the unfamiliar pairs it was 157.5. As these numbers were not normally distributed a non-parametric test of significance was performed, showing a highly significant difference: Wilcoxon test, equivalent z-score = -2.70, p = 0.007.

However the length of contributions in seconds for familiar pairs was very significantly shorter: Wilcoxon equivalent z-score = -2.93, p = 0.003. (In terms of number of tokens per contribution, this difference did not reach significance.)

5.2.3 Pairs with eye-contact produce longer contributions than pairs without. Contribution lengths both in terms of time and number of tokens are longer in the dyads with eye-contact permitted than those without. With eye-contact the median number of tokens per contribution is 3, without eye-contact it is 2. The median length of contribution is 1.12 seconds with eye-contact, and 0.92 without eye-contact. These differences are highly statistically significant. (For number of tokens, Wilcoxon equivalent z-score = -3.33, p = 0.001; for time in seconds, equivalent z-score = -8.30, p < 0.0005.)

Again, this effect can also be seen by tabulating the frequency of single-word contributions: the rate is significantly lower in the group with eye-contact, as shown in Table 6.

Contribution length :	1 word-token	more than 1 word-tokens
Eye-contact allowed	4533	7241
Eye-contact absent	5788	8631

Table 6. Cross-Tabulation of 1-Word Contributions by Eye-Contact.

The rate of single-word contributions is 38.5% in the pairs with eye-contact and 40.1% in those without. Statistically, this difference is highly significant: Chi-squared = 7.25, df = 1, Likelihood ratio = 7.32, p = 0.007.

Some interaction-management functions that can be managed by gaze if eye-contact is present have to be managed by (short) verbal/vocal contributions if it is absent, thus increasing the total of short contributions. As noted by Doherty-Sneddon *et al.* (1997: 113): "speakers attempt to confirm their listeners' understanding or agreement more often when they cannot see one another."

5.2.4 Males produce more tokens per contribution than females (in about the same time). Overall, male speakers produce longer contributions in terms of the number of tokens per contribution, though not in terms of the length of their contributions in seconds. The median number of tokens per contribution is 3 for males and 2 for females. The median length in seconds is 1.03 for males and 1.00 for females. The first of these differences is very highly significant (Wilcoxon test, equivalent z-score = -4.12, p < 0.0005) but the second fails to reach significance (equivalent z-score = -0.237, p = 0.812).

It seems unlikely that males actually speak faster than females so this result might be explained by males speaking with fewer hesitation pauses, though we have not verified this.

Female speakers also produce a higher proportion of single-word contributions than males (40.6% versus 38.3%), as shown in Table 7. This association is very highly significant (Chi-squared = 14.65, df = 1, Likelihood ratio = 14.75, $p < 0.0005$).

Contribution length =	1 word-token	more than 1 word tokens
Female speaker	5188	7593
Male speaker	5133	8279

Table 7. Cross-Tabulation of 1-word utterances by Speaker's Gender.

A plausible reason for this difference is that women tend to give more active-listening signals such as "yeah" than men. "Several American studies have found women providing more backchanneling than men" (Eckert & McConnell-Ginet, 2003: 110).

5.3 Gaps and overlaps

As well as the lengths of contributions, we examined some effects of the lengths of the intervals between contributions (which, when negative, signify overlapping talk).

5.3.1 Overlaps are relatively common. In the corpus as a whole transitions where one speaker begins before the previous speaker has finished comprise 42 percent of all transitions. Table 8 shows the absolute and relative frequencies of overlaps (transitions where the inter-speaker interval is negative), long gaps (where the interval exceeds 1 second) and "normal" intervals (non-negative but less than 1 second).

Gap Type	Frequency	Percentage
Overlap	11011	42.04
"Normal"	12571	47.99
Long	2611	9.97

Table 8. Frequencies of three categories of inter-speaker interval.

This is broadly comparable to the finding of Iwa *et al.* (1998) that 45% of the utterances by participants in the Japanese version of the Map Task overlapped, and would not be surprising except in view of one of the "grossly apparent facts" about conversation listed by Sachs *et al.* (1974: 700): "Transitions (from one turn to a next) with no gap and no overlap are common. Together with transitions characterized by slight gap or slight overlap, they make up the vast majority of transitions". The present results do not directly contradict that assertion because a contribution as defined here is not exactly what conversation analysts mean by a turn, and because the meaning of "slight" and "vast majority" is ill-defined. Nevertheless they suggest that simultaneous speech is far from abnormal.

5.3.2 Average gapsize is longer when eye-contact is allowed. Figure 2 depicts the distributions of mean gapsize (inter-speaker interval) for all 128 conversations, 64 with eye-contact allowed and 64 without eye-contact. It will be seen that when eye-contact is allowed the average size of inter-speaker intervals tends to be longer.

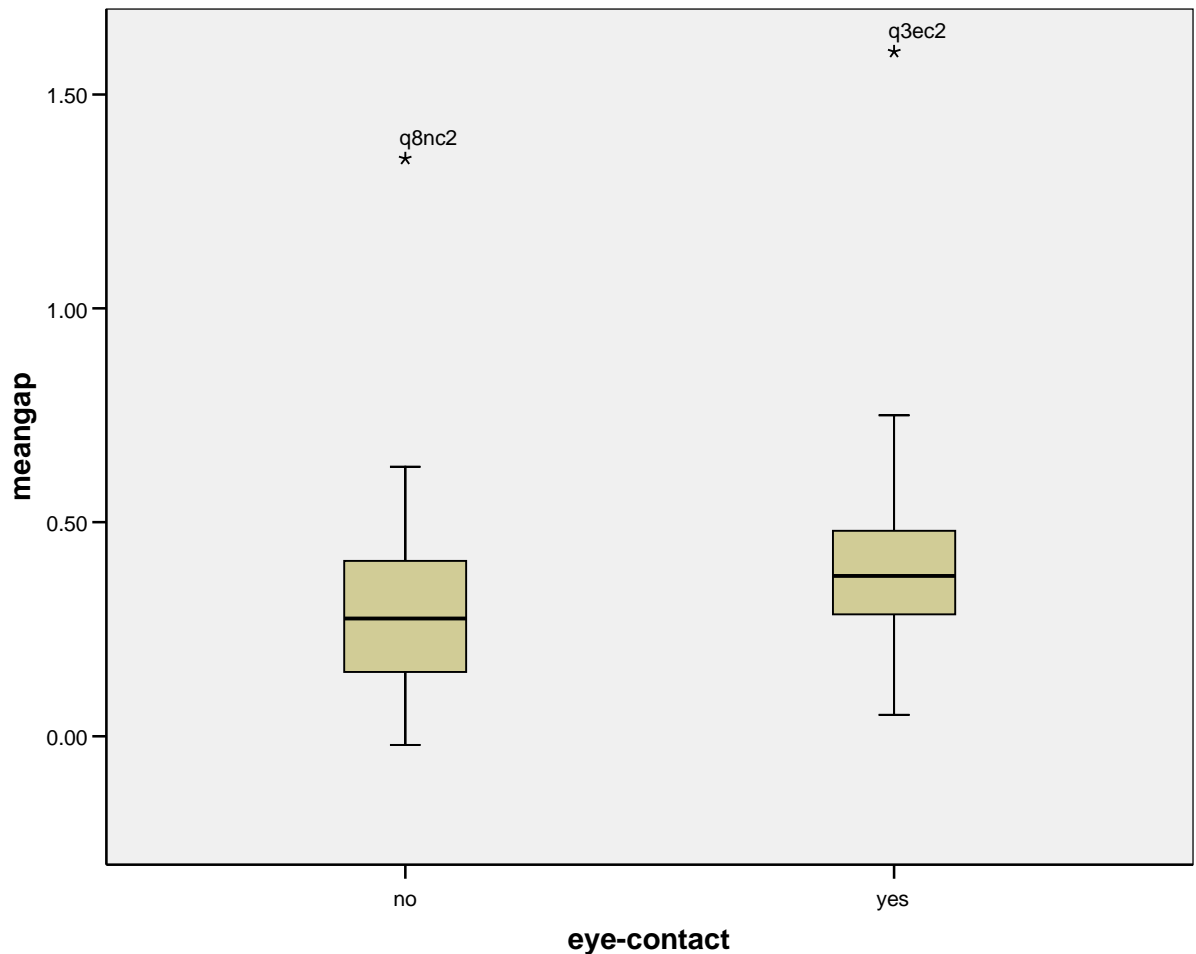


Figure 2. Boxplot of mean inter-speaker interval when eye-contact absent or present.

This difference is highly significant: Wilcoxon equivalent z-score = -2.93, $n=128$, $p = 0.003$. This finding is broadly compatible with the results of Bull and Aylett (1998), although they defined utterances differently.

5.4 Associations with the outcome variable

5.4.1 Longer contributions by followers are associated with worse performance scores. A number of derived variables relating to contribution lengths were computed and correlated with the outcome measure, path deviation score, for each of the 128 dialogues. These were: total number of contributions, total time of talk, mean length of contribution (in tokens and seconds), mean lengths of contributions for giver and follower separately (in tokens and seconds), and the rate of 1-word contributions (overall, for giver and for follower). Since the path deviation score was not normally distributed (Kolmogorov-Smirnov statistic = 1.695, $p = 0.006$) nor were several of the other variables, a non-parametric correlation, Spearman's rho, was used. Only one of these eleven variables was found to correlate significantly with outcome score, namely the mean length of the follower's contributions in seconds ($\rho = 0.264$, $p = 0.003$). This correlation is positive, meaning that longer contributions by the follower (not the giver) are associated with worse results in the task. This may be because long contributions by the follower are a kind of "losing-the-plot" signal. In other words,

when the interaction is going smoothly, the follower won't have anything very complicated to express, resulting in shorter contributions.

5.4.2 In the worst-performing quartile, givers produce fewer 1-word contributions. It is also the case that if the 128 dialogues are partitioned into those with a deviations score of over 100 (the upper, i.e. worse-performing, quartile) and the rest, the givers in the worse-performing quartile have a significantly lower rate of 1-word contributions, 11.12% versus 13.23%, than the rest (t-test: $t = 2.38$, $df = 126$, $p = 0.02$). This may indicate a lesser rate of monitoring the effect of their instructions by the givers in the less-successful group.

5.4.3 Overlaps can be good news! For the 128 dialogues taken as a whole, there are significant negative correlations between the total number of overlaps and the deviation score (Spearman's $\rho = -0.215$, $n=128$, $p = 0.015$) as well as between the proportion of overlaps and the deviation score (Spearman's $\rho = -0.280$, $n=128$, $p = 0.001$). The negative sign implies that more frequent overlapping speech is associated with lower deviation scores, i.e. better task performance. This is another blow for the view that overlapping talk is in some sense pathological. In the words of Tannen (1994: 60): "many instances of overlap are supportive rather than obstructive."

Figure 3 shows that this relationship is non-linear, despite generating a significant rank correlation. This L-shaped distribution might be better characterized as a logical NAND: it is possible to have high deviation scores or a large number of overlaps (or neither) but not both. The reason why the upper-right quadrant of the graph is, in effect, forbidden, is not clear.

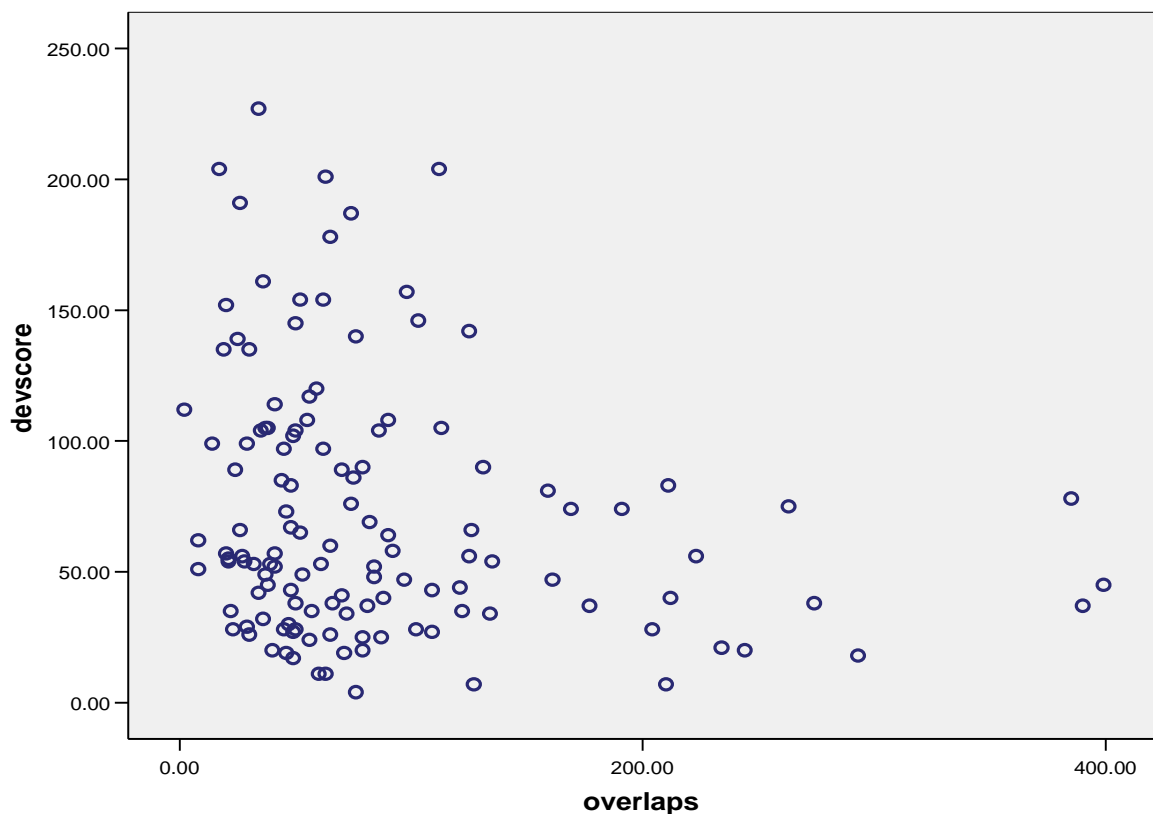


Figure 3. Relationship between number of overlaps and deviation score.

5.4.4 Same-sex pairs perform better on the task than mixed-sex pairs. Another finding of interest is that same-sex pairs tend to have lower (i.e. better) path deviation scores. This difference is significant (Wilcoxon test, equivalent z-score = -2.33, p = 0.02). This effect does not seem to have been noted previously in the literature.

5.5 Vocabulary and positioning

If TST1 were merely chopping up the speech stream at random, one would expect the words next to boundary-points to be a random selection of the words in the corpus, with the same probability of occurring near a boundary as anywhere else. On the other hand, if the technique is slicing the speech-stream at socially or linguistically meaningful junctures, one would expect position in a contribution to have an effect on vocabulary. One way of examining this issue is to look at the initial words of the 26193 contributions.

5.5.1 Vocabulary in initial positions is not a random subset of the overall vocabulary. Table 9 lists the thirteen most common contribution-initial words in the corpus, which between them account for over 53 percent of contribution-initial words, along with their overall frequencies in the corpus, their frequencies as initial words, and their frequencies as initial (and only) words in 1-word contributions.

Word	Whole Corpus		Initial Position		Solo Word	
	Frequency	Percent	Frequency	Percent	Frequency	Percent
right	6932	4.5077	3838	14.6528	1735	16.8104
okay	2458	1.5984	1467	5.6007	889	8.6135
yeah	1710	1.1120	1342	5.1235	783	7.5865
uh-huh	1460	0.9494	1286	4.9097	957	9.2724
and	3464	2.2526	1085	4.1423	103	0.9980
no	1315	0.8551	923	3.5238	390	3.7787
mmhmm	903	0.5872	870	3.3215	763	7.3927
the	12870	8.3691	751	2.8672	145	1.4049
so	1647	1.0710	648	2.4739	140	1.3565
well	1090	0.7088	505	1.9280	85	0.8236
to	4447	2.8918	464	1.7715	112	1.0852
you	4657	3.0284	459	1.7524	90	0.8720
oh	566	0.3681	403	1.5386	95	0.9205
sums =	153780		26193		10321	

Table 9. Most common words in initial position.

It can easily be seen that the relative frequencies of the words in initial positions do not distribute themselves according to the relative frequencies in the corpus as a whole. For example, the token "uh-huh" accounts for less than one percent of the tokens in the corpus, yet comprises almost five percent of the tokens found in initial positions. In the other direction, the commonest word in the corpus, "the", accounts for over eight percent of the word tokens, but only 2.87 percent of those in initial position.

With such obvious discrepancies, a statistical test of significance is hardly necessary, but for completeness a Chi-squared test was conducted on the 13-by-2 matrix containing frequency counts for these words in initial and non-initial positions. As expected this yielded a very highly significant result, Chi-squared = 15547.69, df =

12, log-likelihood = 16769.39, $p < 10^{-300}$. It can be stated with extreme confidence that words in initial positions of contributions are not a random selection of the vocabulary of the corpus as a whole.

5.5.2 Some words in initial positions are very much more likely to stand alone than others. In addition, a further test was performed on the 13-by-2 matrix of counts of these words (all initial) which did or did not have subsequent words following them in the contributions which they started, as shown in Table 10.

Word	Solo	Initial but not Solo (with followers)
right	1735	2103
okay	889	578
yeah	783	559
uh-huh	957	329
and	103	982
no	390	533
mmhmm	763	107
the	145	606
so	140	508
well	85	420
to	112	352
you	90	369
oh	95	308

Table 10. Initial words with and without followers.

Here again the result was very highly significant (Chi-squared = 2670.57, $df = 12$, log-likelihood = 2908.08, $p < 10^{-300}$). This shows that some initial words, such as "mmhmm", tend to stand alone while others, such as "well", tend to inaugurate a contribution of more than a single word.

5.5.3 Words with different functions behave differently with respect to positioning. These results alone would be sufficient to show that the division of speech produced by TST1 is very far from arbitrary, but that is merely a first step. It is perhaps more interesting to take a look at these words individually in respect of their positioning and see what implications emerge about vocabulary choice by the participants. Figure 4 displays each of these word tokens as a point on a two-dimensional graph in which the horizontal axis measures the degree to which that word is preferentially found in initial positions (as compared to its occurrence rate in the whole corpus) and the vertical axis is the degree to which the word is preferentially found as a single-word contribution (as compared to its occurrence rate among words in initial positions).² These words fall into three main groupings.

Continuers	High-frequency words that appear less often in initial positions than expected according to their overall frequency	the, to, you
Initializers	Words relatively frequent in initial positions that usually initiate a multi-word contribution	and, oh, so, well
Solitaires	Words very frequent in initial positions that frequently constitute a single-word contribution in themselves	mmhmm, no, okay, right, uh-huh, yeah

Table 11. Three kinds of initializing words.

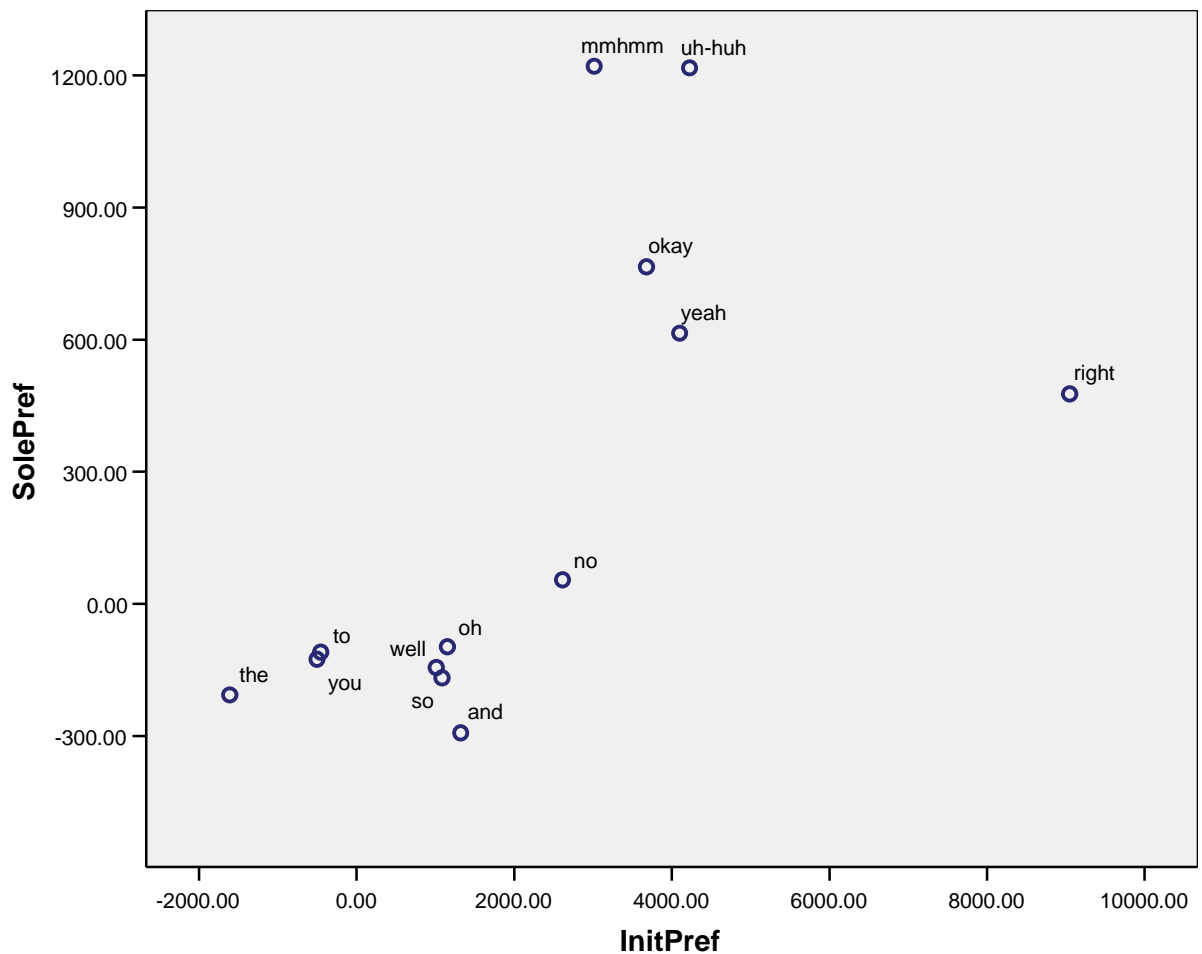


Figure 4. Plot of words in space of Initial versus Solo Preference.

Even within the "solitaires" there are distinguishable subgroups. The word that is by far the commonest in initial position, "right", also has a high rate of appearance as a solo contribution, but it is less likely to appear alone than some words that are less frequent as initials. In this respect it is far outdone by "mmhmm" and "uh-huh",³ which are transcription conventions for the kind of paralinguistic vocal signals of assent or confirmation known as backchannel communications (Yngve, 1970; Schegloff, 1982). As for "no", it is slightly more likely to stand alone when in initial

position than usher in a multi-word contribution, but not by nearly as much as the other tokens in the group labelled "solitaries".

Thus differences of function and meaning manifest themselves as differences in positional preference within the TST1-defined contributions.

6. Discussion

This investigation is modest in scope. Its primary objective has been to introduce a simple method of dividing transcribed speech into segments on a strict chronological basis, and to make an initial test of the utility of that method (TST1) by applying it to a publicly available spoken dataset, the Maptask corpus.

Digital recording, and analysis, of speech is becoming more important in a number of fields, and its importance is likely to grow. Anyone who has been involved in transcribing recorded speech will know that dealing with simultaneous utterances by more than one speaker is one of the most problematic and time-consuming aspects of the process. From a representational point of view, the advantage of TST1 is that it is automatic. It sidesteps the normative considerations inherent in the concept of turn-taking, and merely sorts all the vocables of a conversation into sequence according to their onset-time -- noting by a line-feed and a speaker-prefix when the current vocable has been produced by a different speaker from the last one.

Thus it frees the transcriber from the problem of deciding which speaker "has the floor" and the associated problem of how to deal with cases where a speaker other than the one who is deemed to "have the floor" says something.

Of course this means that the resultant layout does not directly correspond to some established notions of what constitutes turn-taking. Whether this is a serious loss or not depends on the analysts' purpose and cannot be decided generally. However, in practice (at least on one corpus in one language) the speaker-division patterns produced by TST1 do not look outlandish: they are intelligible as dialogue even if somewhat vertically "stretched" by the standards of a typical playscript.

The main question is whether this way of dividing speech into chunks offers any benefits to a researcher studying spoken interaction. We argue that the results in section 5 suffice to show that features derived from the division imposed by TST1 yield indicators that can serve as diagnostic or predictor variables for significant aspects of the interaction. To recapitulate:

- Speaker role is very strongly associated with length of contribution;
- Familiar participants produce more contributions than unfamiliar ones;
- Presence or absence of eye-contact is reflected in average contribution length and average inter-speaker interval;
- Males produce more word-tokens per contribution than females;
- Longer contributions are associated with worse task performance;
- More overlapping speech is associated with better task performance.

In addition, very clear association between positioning within a contribution and vocabulary selection shows that TST1-defined boundaries are linguistically

meaningful. For instance, vocables considered as backchannel signals (such as "mmhmm" and "uh-huh") show strong positional preferences, being very much more common as initial tokens than at other positions and also, given initial position, much more common as sole-word contributions.

The chief disadvantage of this method of talk-division is that it requires accurate timings of each of the vocables uttered by all participants in a conversation. However, as automatic word-segmentation software becomes cheaper and more reliable, this problem should become less severe.

Its two great advantages are (1) it is simple; and (2) it is objective. Most existing methods of talk-division are complex: a turn or utterance is typically defined in terms of a mixture of syntactic, semantic, prosodic and functional characteristics -- requiring human expertise with the costs in terms of time and validation that the exercise of human judgement requires. (*cf.* Ford & Thompson, 1996; Carletta *et al.*, 1997; Bull & Aylett, 1998; Koiso *et al.*, 1998; ten Bosch *et al.*, 2004.) TST1 by contrast requires human judgement only in the identification and timing of word boundaries, an uncontentious, though onerous, task. Once that is done the segmentation of the speech stream is automatic, thus approaching closer to the ideal of "letting the data speak".

Acknowledgements

This work was partly supported by Grant R149230035 of the UK Economic and Social Research Council as part of the DReSS Project sponsored by the National Centre for e-Social Science and partly by the School of Psychology, University of Nottingham. The authors would also like to thank Professor Jean Carletta and Dr Amy Isard, of the Human Communication Research Centre of the University of Edinburgh, for answering questions about the Maptask data and kindly supplying the path deviation scores which are not currently included on the Maptask website. In addition, we would like to thank Professor Claire O'Malley, of the University of Nottingham, for encouraging us to investigate this corpus in the first place.

References

Anderson, A.H., M. Bader, E. Bard, E. Boyle, G. Doherty, S. Garrod, S. Isard, J. Kowtko, J. McAllister, J. Miller, C. Sotillo & H. Thompson. (1991). The HCRC Map Task corpus. *Language & Speech*, 34, 351-360.

[<http://www.hcrc.ed.ac.uk/maptask>]

Atkinson, J.M. & J. Heritage. (1984) eds. *Structures of Social Action: Studies in Conversation Analysis*. Cambridge: Cambridge University Press.

Boden, D. & D.H. Zimmerman. (1991). *Talk and Social Structure*. Cambridge: Polity Press.

Brown, G., A.H. Anderson, G. Yule & R. Shillcock. (1984). *Teaching Talk*. Cambridge: Cambridge University Press.

- Bull, M.C. & M.P. Aylett. (1998). An analysis of the timing of turn-taking in a corpus of goal-oriented dialogue. *Proceedings of ICSLP-98 Sydney, Australia* (4). 1175-1178.
- Carletta, J., A. Isard, S. Isard, J.C. Kowtko, G. Doherty-Sneddon, A.H. Anderson. (1997). The reliability of a dialogue structure coding scheme. *Computational Linguistics*, 23(1), 13-31.
- Carter, R. (2004). *Language and Creativity*. London: Routledge.
- Doherty-Sneddon, G., A.H. Anderson, C.E. O'Malley, S. Langton, S. Garrod & V. Bruce (1997). Face-to-face and video-mediated communication: a comparison of dialogue structure and task performance. *J. Experimental Psychology: Applied*, 3(2), 105-125.
- Eckert, P. & S. McConnell-Ginet (2003). *Language and Gender*. Cambridge: Cambridge University Press.
- Ford, C.E. & S.A. Thompson (1996). Interactional units in conversation: syntactic, intonational, and pragmatic resources for the management of turns. In Ochs, E., E.A. Schegloff & S.A. Thompson (eds). *Interaction and Grammar*. Cambridge: Cambridge University Press.
- Iwa, J., M. Enomoto, K. Ohya, K. Shimano & S. Tutiya. (1998). A study of speech overlap in Map-task dialogues. SIG-SLUD-9802-3, 15-21.
- Koiso, H., Y. Horiuchi, S. Tutiya, A. Ichikawa & Y. Den. (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), 295-321.
- Leech, G., G. Myers & J. Thomas. (1995) eds. *Spoken English on Computer: Transcription, Markup and Applications*. Harlow: Longman.
- MICASE (2004). Data collection, transcription and mark-up. [<http://www.lsa.edu/eli/micase/index.htm>]
- Payne, J. (1995). The COBUILD spoken corpus: transcription conventions. In Leech, G., G. Myers & J. Thomas. (eds) *Spoken English on Computer: Transcription, Markup and Applications*. Harlow: Longman. 203-207.
- Sachs, H., E.A. Schegloff & G. Jefferson. (1974). A simplest systematics for the organization of turn-taking for conversation. *Language*, 54(4), 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*. Washington DC: Georgetown University Press. 71-93.
- Schiffrin, D. (1994). *Approaches to Discourse*. Oxford: Blackwell.

Sinclair, J. (1991). *Corpus, Concordance, Collocation*. Oxford: Oxford University Press.

Tannen, D. (1994). *Gender and Discourse*. Oxford: Oxford University Press.

ten Bosch, L., N. Oostdijk & J.P. de Ruiter. (2004). Durational aspects of turn-taking in spontaneous face-to-face and telephone dialogues. *Proc. 7th Int. Conf. on Text Speech & Dialogue*. Brno, Sept. 2004.

Yngve, V.H. (1970). On getting a word in edgewise. *Papers from the 6th Regional Meeting of the Chicago Linguistic Society*, Chicago: University of Chicago, 567-578.

¹ This is functionally equivalent to the Mann-Whitney test, but seems to have superseded it among statisticians.

² Position on the horizontal dimension is computed as $2 \times O_j \times \ln(O_j/E_j)$ where O_j is observed frequency of token j and E_j is expected frequency. This is the component formula of the log-likelihood statistic, here used as an index. Expected frequencies for the horizontal dimension were calculated from word-frequencies in the whole corpus; expected frequencies in the vertical dimension were calculated from word-frequencies in initial position. Thus vertical height indicates the degree to which a token, given that it is already in initial position, is the sole token in a contribution.

³ In North American English the token "mmhmm" might normally be thought to indicate a positive sign, of agreement or permission to continue speaking, while "uh-huh" would probably be considered a negative signal, e.g. of disagreement or surprise. However, in Scottish English, the sound rendered here as "uh-huh" is most often (though not invariably) another positive or agreeing signal.