A PROFILE OF THE DISCOURSE AND INTONATIONAL STRUCTURES OF ROUTE DESCRIPTIONS

Sandra Williams[†] *and Catherine I. Watson*[‡]
[†]Microsoft Research Institute, [‡]Speech Hearing and Language Research Centre
Macquarie University, Sydney, NSW 2109, Australia
swilliam@mri.mq.edu.au, watson@srsuna.shlrc.mq.edu.au
http://www.mri.mq.edu.au/~swilliam, http://www.shlrc.mq.edu.au/~watson

ABSTRACT

A corpus of spontaneous route descriptions was collected from 8 speakers (5 males and 3 females). The corpus was labelled according to the ToBI standard and a discourse analysis was completed. Four discourse tour acts were identified and these were found to occur mainly in non-embedded linear sequences. In general, the intonation of each route description was characterised by a single intonational phrase containing many intermediate phrases. There was a tendency for boundaries between tour acts and intermediate phrases to coincide, but there is usually more than one intermediate phrase to one tour act.

Keywords: discourse, intonation, prosody, route descriptions.

1. INTRODUCTION

This study draws some preliminary conclusions about relationships between intonational structures and discourse structures in a corpus of spontaneous spoken Australian English. The corpus contains 56 monologue descriptions of routes around a university department building.

In this paper we describe how we collected the corpus and how we carried out a ToBI intonational analysis [1], and a discourse analysis. We then correlate the results of both analyses. The results of this study will be used to improve a Natural Language Generation (NLG) system [2] that generates prosodically annotated routes to be spoken by a speech synthesiser.

A previous study of discourse structure and intonation in a corpus of American English route descriptions [3] by Davis and Hirschberg (D&H) was also used to inform the design of a system for automatically generating synthetic spoken route descriptions. Since we were attempting similar analyses, we have compared the results of our analysis with their findings and found some differences.

2. METHOD

2.1 Corpus Collection

Our corpus consists of spontaneous speech from speakers giving instructions on how to get from the reception desk of a university department to 7 different destinations within the department. The speaker stood facing reception, and gave instructions to the interviewer, who was facing the speaker. The predetermined places of the interviewer and speaker ensured that expressions referring to places, objects and people were not subject to variation due to their changing positions.

The speakers were eight members of the department who were familiar with the building. There were five males who we code-named *ma*, *mb*, *mc*, *md* and *me* and three females who we code-named *fa*, *fb*, and *fc*. All were native speakers of Australian English and aged between 20 and 40 years. The speakers were asked seven questions in total. Five were of the form "How do I get from here to X's office?", where X was the name of a person in the department. The remaining two questions were in the form "How do I get from here to the Y room?", where Y was the name of a room. This gave us a total of 56 route descriptions in our corpus.

Conceptually, routes within a building can be thought of as sequences of *path elements* interleaved with *turns.* This is the underlying formal model of route descriptions we have been using in our computational language generation experiments.

In our computational model of routes, part of which is directed towards providing a mobile robot with appropriate movement instructions, we need to explicitly include both an initial turn into the first direction of movement and the final turn to enter the room being directed to (since doors are in this domain are generally to one or other side of the final path element). As it turns out, human speakers vary as to whether they do or do not include these "orientation" turns.

For example, the two following utterances from our corpus describe route 3:

md: 'Well you turn straight around, walk past the photocopy room to the next corridor, turn left, keep walking along and it's the next door on the left'

fa: 'Go down the corridor behind you, turn left, and it's the first door on the left'

Speaker *md* makes the first orientation turn explicit whilst speaker *fa* leaves this as something that the

hearer must infer from the location 'behind you'. Both speakers leave the final orientation turn to be inferred.

Three of the routes involved three *path elements* and four *turns*, two routes involved two *path elements* and three *turns*, one route involved one *path element* and two *turns*, and the last route involved only one *turn* and *path element*.

Recordings were made directly onto the hard disk of a Dell Latitude laptop PC computer using Microsoft Windows 95 Sound Recorder and recording with a high quality Sony ECM-44B lapel-pin microphone. The speech was sampled at 11.025kHz and quantised as a 16 bit number. The Sound Recorder settings were saved to ensure the same ones were used for every recording.

2.2 Acoustic and Intonation Analysis

To analyse the intonation of the corpus, pitch contours of the recorded speech were calculated with Entropic WAVES+. The data was then marked up according to the ToBI method [1] using the EMU labeller [4]. Start time and end time boundaries of all intonational phrases, intermediate phrases, and words were labelled. Pitch accents were labelled as single points in time. The labelling was carried out by a trained phonetician, and was checked by the authors.

2.3 Discourse Analysis

In our analysis of the route descriptions provided by the subjects, we have chosen to segment these into distinct utterances optionally preceded by cue phrases. Cue phrases are phrases such as 'and then', 'but', 'because' identified according to the rules devised by [5]. Our notion of what counts as an utterance here is clearly influenced by our underlying formal model of routes, but we believe it to be intuitively quite plausible. Thus, in the example spoken by fa, above, we consider there are three distinct utterances, these being:

A: *Go down the corridor behind you*
B: *turn left*
C*: and it's the first door on the left*

We classify these utterances in the following way. At something like the level of a speech act analysis [6], we note that the vast majority of utterances in our corpus fall into three types which we call here *Instruction*, *Confirmation* and *Disclaimer*:

1. *Instruction:* this is an instruction directly concerned with the navigation of the route, e.g. 'turn left', 'walk past the photocopy room'.

2. *Confirmation:* this is a confirming question, asked usually at the beginning of the route description, often repeating the name of the destination. E.g.'Stephen Green's office?'

3. *Disclaimer:* this is a comment on the information being given, e.g. 'but I'm not sure what number' or 'There are two ways you can go'.

There were very few examples of *Confirmation* and *Disclaimer* in our corpus (only four of each). Because these utterances are not directly concerned with the route description itself, we concentrated on those utterances which we have categorised as *Instructions*. For our purposes, finer discrimination amongst these was necessary, and so we adopted the notion of tour acts proposed by D&H.

Our tour acts were divided into four distinct types:

a. *path:* this is an instruction to follow a pathway, examples are: 'Follow this corridor to the end'; 'Walk down the corridor to my right'; 'keep walking until you get to the end'; and so on.

b. *turn:* this is an instruction to make a turning movement. Examples are: 'Turn left'; 'Take the next on the left'; 'Spin round 180 degrees'.

c. *landmark:* this draws attention to some unique feature that will help with navigation. Examples are 'There is a green EXIT sign' and 'There is a little space'.

d. *locate*: this locates the destination: 'His office is the first on the left'; 'Her office is on the other side of the building'. These can occur at the beginning of the route description as a summary, or at the end.

The four types differ from those D&H proposed. The D&H tour acts are directed towards a driver following a highway route. They include 'start', 'stop' and a number of types of road junction turns, they also define a set of route 'cues' to determine when an act should be carried out, e.g. 'When you get to the end of the road, then ...'.

The start and end time boundaries of the tour acts were also labelled using EMU [4].

3. RESULTS

3.1 Results of the Intonation Analysis

The general characteristics of the corpus are demonstrated by the overall frequencies of the tones and accents shown in Table 1 below:

intermediate tones	L- (164), H- (127)
accents	H* (427), L+H* (21), !H* (19), L* (3), L*+H (2)
contour types	L-L% (65), H-L% (12), H-H% (3)

Table 1 Frequencies of ToBI annotations with number of occurrences in parentheses.

Intermediate tones are fairly evenly distributed between L- and H-. The large number of H- is due to the presence of the Australian English high rising terminal [7]. The pitch accents are almost all H* which was expected from the the results of [8] in their analysis of radio news speech. D&H found a frequent use of H*+L accents in their corpus which they claim are

typical of didactic speech. However D&H use an analysis predating ToBI, and in ToBI the H*+L accent was merged with the H* accent. This may account for the differences in the pitch accents found in our study and D&H's study.

The contour types are almost all L-L% but there is a noticeable number of H-L%. The L-L% tone is associated with neutral declaratives and the H-L% tone can be associated with a slightly impatient recital of lists [9]. Both these boundary tones are therefore entirely consistent with what we would expect in route descriptions. The exceptions are two confirming questions with the analyses L* H-H% and H* H-H%, and one cue phrase "okay" with the analysis L* H-H%.

As expected, on investigating the sequence of ToBI labels, there was no particular tune to characterise each tour act. The only measurable feature was the expected overall declination of the f0 contour over the entire route description.

3.2 Results of the Discourse Analysis

The first row in Table 2 shows the total frequencies in the corpus of the tour acts described above. The majority of these are *path*, *locate* and *turn* tour acts, with approximately equal numbers of each. There are relatively few occurrences of *landmark*.

The second row shows the most frequently occurring tour acts at the beginning of the route descriptions. Notice that *path* and turn begin the description, as in the examples in section 2, roughly the same number of times. When the locate tour act begins a route description, it may end the description too. As one would expect, initial *locates* describe more generally where the destination is, whilst final *locates* are more specific. For instance initially: 'It's in the far corner of the building' and finally: 'It's the first on the left'. Possibly a finer distinction between these two types could be drawn here that would be useful for NLG.

The third row of Table 2 shows the vast majority, 50 of a total of 56 routes, end with a *locate* tour act. This is unsurprising since it is normal to indicate to the hearer that the destination has been reached with this utterance.

Unlike D&H, we found few cases of embedded utterances in the discourse structures of our corpus. Where utterances were linked by cue phrases, by far the majority of these were the simple sequential links one would expect such as 'and' (57 occurrences), and 'and then' (23 occurrences). Most route descriptions in our corpus are simple linear sequence of utterances. Furthermore the trigram frequencies (i.e. frequencies of occurrence of tour act B when it is preceded by tour act A and followed by tour act C) shown in the final row of Table 2 confirm these sequences of utterances conform to our formal computer model described in Section 2.

tour act frequencies in corpus	path (61), locate (63), turn (60), landmark (5)
first tour act in description	path (18), turn (17), locate (12)
last tour act in description	locate (50)
tour act trigram sequences	path turn locate (20), turn path turn (13), path turn path (11),

Table 2. Tour act occurrences (actual numbers in brackets)

3.3 Correlations between ToBI Phrases and Tour Acts

Table 3 shows for each speaker and each route the number of ToBI intonational phrases (I) and intermediate phrases (Im) and tour acts (TA) for each route description. In the majority of the 56 route descriptions analysed in this study, the entire route description was spoken in a single ToBI intonational phrase. Only 18 descriptions were spoken in more than one intonational phrase. This was an unexpected finding since D&H appear to relate one intonational phrase to one tour act.

Of the 50 route descriptions ending in a *locate* tour act, 47 were labelled L-L% at the end of the phrase, as we expected. However, only 29 of these coincided exactly with the final L-L% phrase, others had additional phrases at the beginning.

As we expected, different speakers use different numbers of tour acts and intermediate phrases to describe the same route. Table 3 also shows the total number of turns and paths as characterised by our formal computational model. It can be seen that routes with few turns and paths (Routes 5 and 7) also have fewer tour acts and intermediate phrases, but there is no particular pattern present.

Most route descriptions are made up of several tour acts and several intermediate phrases. The average number of intermediate phrases per route description was five. A closer inspection of the temporal boundaries of the 296 intermediate Phrases, and the 219 tour acts reveals that 131 of them had exactly the same boundaries, 196 began in the same place, and 198 had the same end times. There are more intermediate phrases than tour acts, but most tour acts are made up of one or more intermediate phrases. D&H appear to find a one-to-one relationship between tour acts and intonation phrases. Our analyses show a relationship between tour acts and ToBI phrases at the intermediate phrase level, not the intonational phrase level, and the relationship is not one-to-one. D&H used a version of intonational analysis that predates ToBI, which may account for some differences in analysis, and there are differences in individual human labellers. However this appears to be a unique feature of our corpus.

spkr	Route 1			Route 2			Route 3			Route 4			Route 5			Route 6			Route 7		
	3 turns, 2 paths			4 turns, 3 paths			4 turns, 3 paths			4 turns, 3 paths			1 turn, 1 path			3 turns, 2 paths			2 turns, 1 path		
	I	Im	TA	I	Im	TA	I	Im	TA	I	Im	TA	I	Im	TA	I	Im	TA	I	Im	TA
fa	1	3	3	1	3	3	2	3	3	1	3	5	1	2	1	1	6	3	1	3	1
fb	1	6	4	1	6	4	1	6	5	1	6	5	1	3	3	1	6	4	1	4	3
fc	3	5	4	2	4	4	2	4	4	1	7	5	1	3	4	1	5	4	1	3	3
ma	1	7	4	1	1	1	1	11	8	3	11	5	1	7	3	2	5	3	1	8	3
mb	5	8	7	3	4	4	3	6	5	1	8	6	1	4	4	1	6	4	1	6	6
mc	2	5	3	3	8	4	1	9	3	1	9	8	1	4	1	1	7	5	2	6	4
md	2	6	4	1	6	3	1	6	5	3	11	12	1	2	3	1	5	4	1	2	1
me	2	6	5	1	5	3	1	5	3	1	6	4	2	3	1	1	5	3	2	3	2

Table 3. Nos. of Intonational phrases (I), Intermediate phrases (Im) and Tour Acts (TA) for each speaker and route

3.4 Correlations between the Pitch Accents and the Discourse Analysis.

We found no correlations here that have not been found before in other studies. As we expected, directional information is very important in route descriptions so it is accented, as demonstrated by the high frequencies of accents on lexical items such as 'left', (56 accented and 9 deaccented), and 'right', (47 accented and 5 deaccented). Also information that distinguishes one building entity (e.g. office or room) from others in the vicinity tends to be accented. For instance 'end' is accented in 'the office at the end', and 'photocopy' in 'photocopy room'.

4. CONCLUSIONS

This has been a fairly detailed, but small-scale, study. It provides a unique preliminary profile of the prosodic and discourse structure of spontaneous Australian English speech where there is a scarcity of such studies.

We have identified a set of tour acts to fit the route description data and found that these occur characteristically in non-embedded linear sequences with a final *locate* act. The intonation of the route descriptions is characterised by a single intonational phrase containing many intermediate phrases. There is a tendency for boundaries between tour acts and intermediate phrases to coincide, but there is usually more than one intermediate phrase to one tour act.

It should be noted that no assessment of the quality of the route descriptions was made here. We noted that the 8 speakers varied considerably in their descriptions, but it is outside the scope of this study to say which factors contribute towards a 'good' route description. Although we will experiment with implementing some of the above characteristics in our NLG/speech synthesis system, further investigations are necessary to identify which work best.

5. ACKNOWLEDGEMENTS

Our thanks to Cécile Pereira who labelled the corpus with ToBI annotations and to Robert Dale and Jonathan Harrington for their helpful comments and suggestions.

6. REFERENCES

[1] Silverman, K., Beckman, M., Petrelli, J., Ostendorf, M., Wrightman, C., Price, P. Pierrehumbert, J. and Hirschberg, J. [1992] ToBI: a standard for labelling English prosody. ICSLP92, vol. 2, pp. 867-870.

[2] Williams, S. [1998] Generating Pitch Accents in a Concept-to-Speech System Using a Knowledge Base. ICSLP98, vol. 4, pp. 1159-1162.

[3] Davis, J and Hirschberg, J [1988] Assigning Intonational Features in Synthesized Spoken Directions. ACL88 pp. 187-193.

[4] Cassidy, S and Harrington, J. " EMU: an enhanced hierarchical speech data management system", In *Proceedings of the 6th International Conference on Speech Science and Technology*, Adelaide, 361-366, 1996

[5] Knott, A. and Dale, R. [1994] Using Linguistic Phenomena to Motivate a Set of Rhetorical Relations. Technical Report RP-34 HCRC, University of Edinburgh.

[6] Searle, J.R. [1969] *Speech Acts.* Cambridge University Press

[7] Guy, G. and Vonwiller, J. "The high rising tone in Australian English", In Collins, P. & Blair, D. (Eds) "*Australian English: the language of a new society*",. St Lucia:University of Queensland Press, 1989

[8] Hiyakumoto, L. Prevost, S and Cassell, J [1997] Semantic and Discourse Information for Text-to-Speech Intonation. *Proceedings of the ACL Workshop on Concept to Speech Generation Systems*, Madrid, Spain. pp. 47-56.

[9] Harrington, J. and Cassidy, S. *"Techniques in Speech Acoustics*", Kluwer, in press