MODELLING INTONATION IN DIALOGUE

G. Ayers, G. Bruce**, B. Granström***, K. Gustafson***, M. Horne**, D. House**, and P. Touati ** (Names in alphabetic order)*
**Dept. of Linguistics , Ohio State University, 222 Oxley Hall, 1712 Neil Avenue , Columbus, OH 43210-1298, U.S.A.*
***Dept. of Linguistics and Phonetics, Helgonabacken 12, S-22362 Lund, Sweden*
****Dept. of Speech Communication and Music Acoustics, KTH, Box 70014, S-100 44, Stockholm, Sweden*

ABSTRACT

The analysis of spontaneous dialogue in Swedish is discussed. The methodology, speech material and analysis types are presented, as well as the intonational aspects under study. In particular, tonal downtrend is examined in relation to lexical semantic aspects of topic structure, and the environments in which downstepping occurs are outlined.

BACKGROUND

The framework for our present research on prosody is the project Prosodic Segmentation and Structuring of Dialogue (proZodiag) centered around the description of Swedish. The project is supported within the second phase of the Swedish Language Technology Programme (1993-1996) and represents cooperation between Phonetics at Lund and Speech Communication at KTH, Stockholm [1].

The object of study is the prosody of spontaneous speech and dialogue. The main goals of this research are to increase our understanding of prosody in its natural environment - dialogue and spontaneous speech - and eventually to create a more powerful prosody model.

The background for our present research effort is experience from two decades of study of the prosody of prepared speech in a laboratory setting. The intention of using such laboratory speech has not been to study the prosody of reading, but rather to simulate natural, spontaneous speech. Laboratory speech provides us with a high degree of experimental control but presents an artificial situation with often pragmatically anomalous speech material where the informant's skill in acting becomes important in the recording of the data. For a discussion see [2, 3].

A reasonable question to ask in this context is then why up until fairly recently have there been so few studies on the prosody of spontaneous speech. The apparent reason for this state of affairs is the well testified complexity of prosody; the control of variables is much easier in the phonetics laboratory.

METHODOLOGY

We are exploiting a number of different kinds of speech material: true spontaneous dialogue (recorded with no intention of being studied phonetically), spontaneous 'lab speech' dialogue (with partly predetermined conversational topics but otherwise spontaneous), acted dialogue from scripts (using the 'lab speech' dialogues), dialogue simulation in text-to-speech synthesis, and man – machine dialogue.

For our study of dialogue prosody and spontaneous speech we are using the following types of analysis: analysis of the dialogue structure itself (without taking prosodic information into account), and a two part prosodic analysis: an auditory analysis in the prosodic transcription stage and an acoustic-phonetic analysis which takes F0 and waveform information into consideration.

We conceive of the analysis of the dialogue structure as comprising at least the following aspects: textual aspects (both the structure of conversational topics and its relation to lexical-semantic aspects of focus), turn regulating aspects (keeping, yielding, taking the turn), initiative / response structure, feedback (both giving and seeking feedback), and rhetoric activity (which appears to occur in all kinds of speech, to varying degrees).

In our prosodic analysis of Swedish, focus is on intonational aspects. Our auditory analysis results in a prosodic transcription involving two levels of prominence (accented, focussed), tonal junctures and two levels of grouping (minor, major phrase). The choice of word accent (accent I / accent II) in Swedish is lexically determined. The prosodic categories used are the following:

Tonal Structure		
accented	accent I (HL*)	
	accent II (H*L)	
focussed	accent I ([H]L*H)	
	accent II (H*LH)	
	compound (H*L...L*H)	
juncture	initial (%L; %H)	
	terminal (L%; LH%)	
Grouping		
boundary	minor	|
	major	‖

The acoustic-phonetic analysis is based on F0 and waveform information, whereby both global features such as, for example, F0 level and F0 range and local features such as direction and timing of F0 events are taken into consideration and interpreted in our current prosody model.

The three types of analysis – analysis of dialogue structure, auditory analysis, acoustic-phonetic analysis – involving both symbol and signal information are combined and synchronized with each other in the same ESPS/Waves+ environment. The labelling used (symbol information) is similar to the ToBI transcription system for English [4]. It consists of an orthographic tier (marking the end of words), a tonal tier (symbols of tonal structure), a boundary tier (symbols of grouping), dialogue structure tier (hierarchy of conversational topics), and a miscellaneous tier (with extralinguistic and other information).

An important part of our research methodology is the use of speech synthesis. We are using two different approaches in our research. The first method is a synthesis matching technique to verify the prosodic transcription. Our model is implemented in the ESPS/Waves+ environment, and the input is the prosodic transcription with information about type and time location of tonal turning points. This information (with little segmentation indicated) together with phonetic rules from our prosody model are fed into a modified version of the ESPS/Waves+ synthesizer. The model contour is synthesized and compared to the original. Deviations are then studied, which leads to improvements in the model.

The other approach is to use the KTH text-to-speech synthesis system. Using an experimental version of this system which includes an extended set of prosodic markers, we have a flexible tool for manipulating prosodic parameters. It is particularly suited for testing our hypotheses about prosody on new speech material, specifically the simulation of dialogues.

TONAL DOWNTRENDS

One aspect of our modelling of dialogue intonation involves analysis of intonational downtrends. See for example [5, 6]. According to our earlier study of Standard Swedish in controlled laboratory experiments, the occurrence of a focal accent is a pivot for tonal downstepping [7]. Before a focal accent, non-focal accents do not appear downstepped, while after focus downstepping of successive accents within the same phrase / utterance is a characteristic feature. Thus an early focus in an utterance will typically trigger downstepping, while a late focus will tend to arrest this tonal downtrend. That is to say, downstepped accents occur on information that is given in the context as in (1b) which constitutes an answer to the question in (1a). The underlined words in (1b), which are contextually given in the preceding question, are characterized by downstepped accents. Words in bold print are focussed:

(1) a. Vem lämnar ungen nallar? 'Who gives the kid teddy bears?'
b. **Mamman** <u>lämnar ungen nallar</u> 'The **mamma** <u>gives the kid teddy bears</u>'

In our earlier study of spontaneous speech within the 'KIPROS' project, this regularity was found to appear in spontaneous dialogue as well [8, 9]. Several, typical examples of downstepped pitch patterns were observed which seemed to be triggered by the placement of focal accent in the same way as described above.

In our current analysis of spontaneous speech, downstep has also been seen to correlate with contextually given information. However, it has also been observed to correlate with information that cannot be classified as given. The following examples are taken from our analysis of a dialogue recorded from a Swedish radio program about jazz music called 'What's cooking?'.

In (2), the clause after the word *macka* 'sandwich' does not contain given information in the same sense as in example (1). Yet, it is characterized by downstep in the same way as the underlined words in (1).

(2) Jag har ett litet recept på en varm **macka** | som jag faktiskt har har utvecklats utvecklat en aning
'I've got a recipe for a hot **sandwich** | which I in fact have have improved improved a little'

In this example and in the ones that follow, the underlined words are characterized by downstepped accents, and words in bold print are focussed. Phrase boundaries are also indicated.

It should also be noted that in the examples there is no focal accent present in the underlined phrase itself as a direct trigger of the downstepping. Instead, each successive non-focal accent within the phrase is downstepped.

Thus, the generalization that it is given information that gets downstepped is perhaps too narrow to cover all cases of the phenomenon in spontaneous speech. It would, however, be insightful if one could relate examples like those in (1) and (2) to some more general discourse/semantic parameter(s).

One idea which we would like to pursue in this respect is to relate downstepped information to the development of discourse topics. In this regard, one could say that the downstepped information in (2) is similar to the nonfocal material in (1), in that it can be considered as information which is not central to the development of the topic. By 'central to the development of the topic', we then mean related to the specification of the lexically important/'generic' (see below) referents in the semantic field under discussion as well as the specification of the relationships among these referents. In the specific dialogue under consideration, the central topic involves the description of the ingredients in a recipe for a hot tuna fish sandwich. Just as the downstepped information in (1), being contextually given, does not provide anything new as regards the relationships between the referents *ungen* 'the kid' and *nallar* 'teddy bears', the downstepped information in (2) likewise does not lead to the development of the central topic in the discourse from which it is extracted, i.e. to the description of the discourse referent *macka* 'sandwich'. In other words, there is no information regarding the referents that are relevant to the description of the sandwich. The downstepped material constitutes parenthetical information which is unrelated to the specification of the ingredients in the sandwich, i.e. is non-central to the development of the topic.

Another example of the use of downstep is given in (3):

(3) ...man har vitt **bröd** förslagsvis | och på detta lägger man en röra av.....
'...you take white **bread** for example | and on that you put a mishmash of...'

In this case, noncentral can be related to the level of specificity of the discourse referents. The word *röra* 'mishmash' is semantically nonspecific or 'non-generic' as regards its status in terms of lexical hierarchies. That is to say, it is not "at the level of ordinary everyday names for things and creatures" [10] as regards its relation to the other referents which are mentioned in the dialogue, e.g. *bröd* and the ingredients that make up the 'mishmash': *tonfisk* 'tuna fish', *majonnäs* 'mayonnaise', etc. Thus, one can hypothesize that the downstepping in the utterance in (3) is related to the nonspecificity/nongenericness of the referent mentioned. In other words, one can hypothesize that referents that are central to the topic are ones that are relatively more 'generic' or more 'basic'.

A similar situation can also lead to prosodic downtoning, i.e. when the speaker 'comments on' /specifies an already introduced 'generic' discourse referent as in (4-5):

(4) ...man måste ha en burk **tonfisk**, en burk **crème fraiche** | tonfisk med vatten bara, eh, är bra...
'...you have to have a can of **tuna fish**, a package of **crème fraiche** | just tuna fish in water, uh, is good...'

(5)...en tredjedels burk **majonnäs** || gärna lätt majonnäs där också för att crème fraiche i...
'...a third of a jar of **mayonnaise** || preferably light mayonnaise there too since (there's) crème fraiche in (it)...'

Here the downstepping characterizing the second occurrences of *tuna fish* and *mayonnaise* and their respective specifications can be interpreted as reflecting the noncentrality of that information for the development of the topic, i.e. the central referents (generic terms), *tuna fish*, *mayonnaise* have already been mentioned; the comments concerning the fact that it is good if it is tuna fish in water, and that the mayonnaise should preferably be light, although new information, are relatively unimportant as regards the development of the central theme.

CONCLUSION

Since our material is restricted, we cannot be sure of the generality of the observations made here concerning downstepping. Nevertheless, by pinpointing the environments in terms of the topic structure and related lexical semantic correlates, we can test these hypotheses against more extensive data in future studies.

ACKNOWLEDGEMENTS

This work was carried out under a contract from the Swedish Language Technology Programme (HSFR-NUTEK). We would like to acknowledge assistance from Marcus Filipsson and Birgitta Lastow in developing the analysis by synthesis system.

REFERENCES

[1] Bruce, G., Granström, B., Gustafson, K., House, D. and Touati, P. (1994), "Modelling Swedish prosody in a dialogue framework", *Proc. ICSLP 94*, pp. 1099-1102, Yokohama, Japan.

[2] Beckman, M. (1995), "A typology of spontaneous speech", *Proc. ATR International Workshop on Computational Modeling of Prosody for Spontaneous Speech Processing*, pp. 2.23-2.34, Kyoto.

[3] Touati, P. (1995), "Pitch range and register in French political speech", *Proceedings of the ICPhS -95*, Stockholm.

[4] Pitrelli, J., Beckman, M., and Hirchberg, J. (1994), "Evaluation of prosodic transcription labeling reliability in the ToBI framework", *Proc. ICSLP 94*, pp. 123-126, Yokohama, Japan.

[5] Pierrehumbert, J. and Beckman, M. (1988), *Japanese tone structure*, Cambridge, MA: The MIT Press.

[6] Grønnum, N. (1992), *The groundworks of Danish intonation*, University of Copenhagen: Museum Tusculanum Press.

[7] Bruce, G. (1982), "Developing the Swedish intonation model", *Working Papers*, 22 (Dept. of Ling. Lund Univ.), pp. 51-116.

[8] Bruce, G., Touati, P., Botinis, A., and Willstedt, U. (1988), "Preliminary report from the KIPROS project", *Working Papers*, 33 (Dept. of Ling., Lund Univ.), pp. 23-50.

[9] Bruce, G. and Touati, P. (1992), "On the analysis of prosody in spontaneous speech with exemplification from Swedish and French", *Speech Communication* 11, pp. 453-458.

[10] Cruse, D. A. (1986), *Lexical semantics*. Cambridge: Cambridge U.P.