5aSC2

Poster presented at the 134th ASA Meeting, San Diego, California, USA, 1–5 December 1997.
Abstract published in *JASA*, Vol. 102, No. 5, Pt. 2, November 1997, pp. 3202–3203.

Interaction between prosody and discourse structure in a simulated man-machine dialogue

Robert Eklund
Robert.H.Eklund@telia.se

Telia Research AB
Spoken Language Processing
S-136 80 Haninge, Sweden

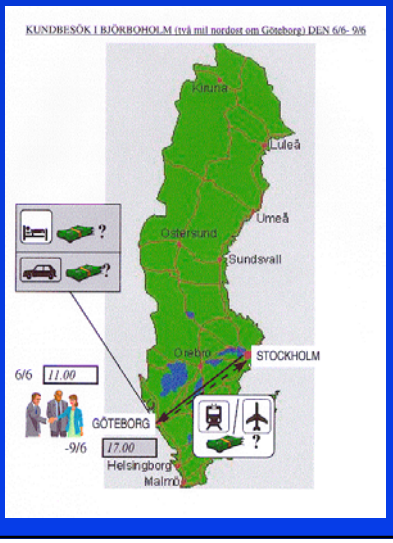
Abstract

Automatic speech understanding systems are beginning to attain a level of sophistication where commercial applications are within reach. However, if humans and machines are ever going to communicate in a natural way, it is of vital importance that language modelling go beyond the sentence level. A profound understanding of discourse structure is required, and to this end, knowledge concerning how prosody interacts with other linguistic phenomena is needed. Not only will better prosodic modelling of discourse lead to better speech recognition/understanding, it will also yield more natural-sounding speech synthesis. This paper reports on a dialogue/prosody project at Telia Research, Sweden. A Wizard-of-Oz simulation of a computerized reservation system was used to collect realistic speech data [pp. 2-- 3]. Fifty subjects were given three tasks each that entailed the reservation of flights, trains, car hire and hotel reservations. To avoid linguistic influence on the subjects' utterances, the tasks were given as maps and icons. A ToBI-style analysis was applied [p. 4], adapted to meet language-specific requirements [pp. 5--6]. The dialogues were analyzed with regard to phrase boundaries, tones, disfluencies, syntax (functions/categories), new vs. given information and pitch range. This paper describes our observations concerning the interaction between prosodic, syntactic and higher-level linguistic phenomena, such as discourse structure [OHs 9, 10, 11].

Data Collection: Wizard-of-Oz Simulation (1)

Method Description

- Data was collected by means of a Wizard-of-Oz simulation of a computerised booking service.
- Subjects were asked to do business trip bookings including flights, trains, hotels, rental cars, taxis etc.
- The subjects were recorded hifi in all the dialogues. In 50 % of the dialogues the wizard was recorded hifi, in the other 50 %, telephone-quality recordings were made.
- Barge-in was not allowed.
- The simulated system engaged in dialogue repair in case of breakdown.
- The subjects were all Telia employees with no experience of speech technology. They had all booked business trips before.
- Subjects were given the tasks as a map of Sweden with dates, times, different means of transport given in iconic, non-verbal form to prevent linguistic bias. The icons were varied between the tasks (see right-hand side of page).
- Some ambiguity was left to enhance linguistic variation.
- For a more detailed description, see MacDermid & Eklund (1997).



Data Collection: Wizard-of-Oz Simulation (2)

Data Description

- 50 subjects were given three tasks each. (Three recording sessions could not be used for technical reasons, leaving 47 successfully digitized sessions.)
- 31 subjects were male; 16 were female.
- Regional and foreign accents were allowed.
- Total number of dialogues: 131
- Total number of utterances: 3602
- Average number of utterances per dialogue: 26.43
  Smallest number: 8
  Highest number: 92
- Average number of words per utterance: 8.11
  Smallest average number: 1.9
  Highest average number: 17.72
- Orthographic transcriptions of all dialoges were carried out, covering both the subjects and the wizard.

Dialogue Sample/Excerpt

TTS Simulated Booking System
FP Subject

TTS: *Välkommen till resebokningstjänsten. Hur kan jag hjälpa dig?* (Welcome to the Travel Booking Service. How may I help you?)

FP: *Jag vill resa från Stockholm till Östersund och sen vill jag resa från Östersund till Luleå och sedan tillbaka till Stockholm.* (I want to go from Stockholm to Östersund and then I want to go from Östersund to Luleå, and then back to Stockholm.)

TTS: *Var vänlig vänta. Vilket datum vill du resa till Östersund?* (Please hold. What date do you want to go to Östersund?)

FP: *Jag vill resa till Östersund den trettionde juni.*
(I want to go to Östersund on the thirtieth of June.)

TTS: *Vilken tid vill du resa?*
(At what time do you want to travel?)

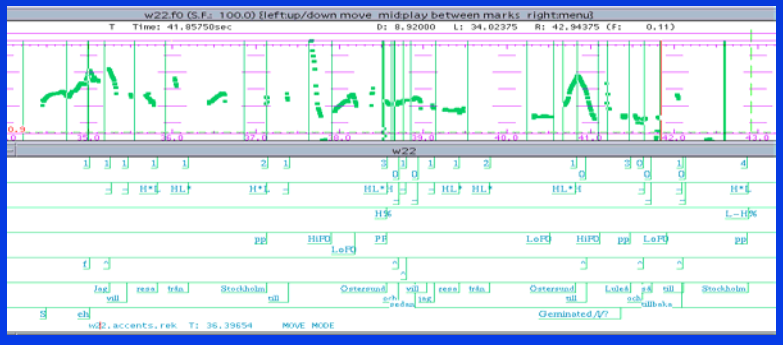
FP: *Klockan femton noll noll.*
(Three o´clock in the afternoon.)

...

Swedish ToBI (1)

Introduction

- The digitized dialogue files are labelled according to the ToBI-style method (Silverman et al. 1992; Beckman & Ayers 1997), adapted to meet Swedish-specific requirements. (Example given below.)
- Swedish-specific labels are adapted to the labels used in the Lund Model project (Bruce & Touati 1990; Bruce & Touati 1992; Bruce & Granström 1993; Bruce et al. 1994; Bruce 1994; Bruce et al. 1995; Bruce et al. 1990).
- The disfluency labels are based on the labels described in Shriberg (1994) and Ostendorf et al. (1997)



Swedish ToBI (2)

Tier 1: Break Index Tier

0	Reductions, contractions and simile.
1	Normal interword spaces in fluent speech.
2	"Trash Bag", contradictory cues.
3	"Intermediate Intonational Phrase".
4	"Full intonational Phrase".
-	Uncertainty diacritic. E.g. 4- indicates.

Tier 2: Accent Tier

HL*	Accent 1, "normal" stress.
H*L	Accent 2, "normal" stress.
H*L ... L*H	Compound word.
HL*H	Accent 1, focused.
H*LH	Accent 2, focused.
-	Deaccentuated word.
!	Downstep (applies to all above)..
-?	Uncertainty whether or not a syllable is accented.
*?	Uncertainty whether or not a syllable is stressed.
X*?	Syllable is stressed, but uncertain how.
(...)	Parentheses around accents that are perceived, but not seen.

Tier 3: Phrase Boundary Tier

%H	Phrase initial high.
%L	Phrase final low.
H%	Phrase final high.
L%	Phrase final low.
L-H%	Phrase final continuation rise.
H-H%	Phrase final high (yes-no question)..
L-L%	Phrase final low.
H-L%	Phrase final, common, ending.
H-L-H%	Phrase final ending.
L-H-L%	Phrase final ending.
?%	Uncertainty whether or not a phrase boundary is present.
X%	Uncertainty what phrase boundary is present.

Tier 4: Peak and Prominence Tier

PP	Perceptually most prominent syllable within each 3/4 phrase.
pp	Perceptually prominent syllables within each 2/3/4 phrase.
HiF0	Visually highest peak(s) within each 3/4 phrase.
LoF0	Visually observable valleys within each 3/4 phrase.

Swedish ToBI (3)

Tier 5: Disfluency Tier

r	Repeated item.
u	Unfilled pause (i..e., silence)
f	Filled pause (e.g., *eehhh* sounds).
e	Elongation (e.g., *biljettennnnn...).*
/	Word fragment.
~	Mispronunciation.
^	Reduced word.

(... etc.)

Tier 6: Orthographic Tier

Normal, lexical, orthography. Reductions and the like are not expressed in the orthographic tier, but in the Break Index tier (with 0) and in the disfluency tier (with ^).

Tier 7: Miscellaneous/Comment Tier

fry	Predefined option.
breath	Predefined option.
asp	Predefined option.
hawk	Predefined option.
cough	Predefined option.
creak	Predefined option.
LLO	Loud Lip Opening.
inh	Inhalation.
exh	Exhalation.
S	(Re)Start, after interruption.
WPS	Words Pronounced Separately.
HYP	Hyperarticulated.

Other Analyses (not in tier form)

The dialogues are also analysed / labelled with regard to:

- Lexical categories.
- Syntactic categories.
- Syntactic functions.
- Open/closed word classes.
- New/Given information.
- Pitch range.

Labeller Consensus Analyses (1)

Introduction

- The ToBI framework has been shown to yield a high degree of agreement between labellers (Pitrelli et al. 1994).
- The current data have so far been labelled by three labellers:

Labeller 1
Active in label decisions. Knowledge about intonational phonology. Has fully labelled dialogues for 24 subjects.

Labeller 2
Active in label decisions Knowledge about intonational phonology. Has labelled dialogues for 2 subjects.

Labeller 3
Not active in label decisions. Little or no knowledge about intonational phonology. Has labelled one dialogue.

Break Index Agreement

All Word Classes

L1/L2:	85 %
L1/L3:	82 %
L2/L3:	77 %
L1/L2/L3:	56 %

Open/Closed Word Classes

L1/L2:	Open:	98 %
	Closed:	77%
L1/L3:	Open:	87 %
	Closed:	80 %
L2/L3:	Open:	85 %
	Closed:	72 %
L1/L2/L3:	Open:	39 %
	Closed:	68 %

Labeller Consensus Analyses (2)

Stress Level Agreement

All Word Classes

L1/L2:	82 %
L1/L3:	64 %
L2/L3:	72 %
L1/L2/L3:	43 %

Open/Closed Word Classes

L1/L2:	Open:	98 %
	Closed:	77%
L1/L3:	Open:	63 %
	Closed:	78 %
L2/L3:	Open:	61 %
	Closed:	67 %
L1/L2/L3:	Open:	20 %
	Closed:	58 %

Stress Levels; New/Given Items

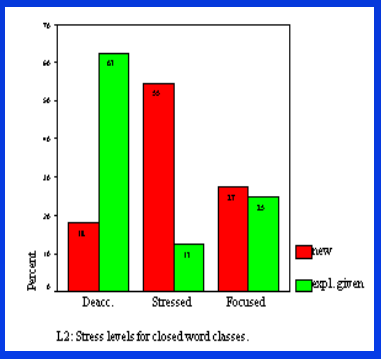
L1/L2:	New:	87 %
	Expl. Given:	74 %
	Impl. Given:	100 %
L1/L3:	New:	74 %
	Expl. Given:	65 %
	Impl. Given:	75 %
L2/L3:	New:	70 %
	Expl. Given:	57 %
	Impl. Given:	75 %
L1/L2/L3:	New:	22 %
	Expl. Given:	30 %
	Impl. Given:	50 %

Preliminary Results (1)

Stress Levels for New/Given Items

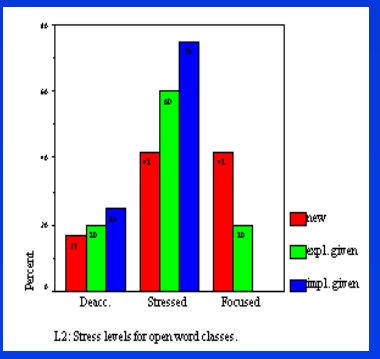
- Sample of one correlation analysis.

Labeller 2: Closed Word Classes



L2: Stress levels for closed word classes.

Labeller 2: Open Word Classes



L2: Stress levels for open word classes.

Preliminary Results (2)

Discourse Markers

- In general, the discourse structure is flat, i.e., embedded topics are not found. One issue/topic is discussed until resolved/closed, whereupon a new issue/topics is opened.
- An array of topic shift markers can be discerned (cf. MacDermid & Eklund 1997).

Closing Discourse Markers

bra [*dâ*]	(Good [then])
okej (*ok*)	(OK)

Opening Discourse Markers

Sedan/sen	(Then)
[*bra*] *dâ*	([good] then)
det är bra	(That's fine)
det blir bra	(That will be fine)
också	(Too, also)
(*sâ*)	(Then)

Syntax

- Mainly declarative sentence structures.
- Swedish allows most items to be fronted, but only a few instances of fronting are found (c. 4 or 5), in connection with misunderstanding/repetitions.
- Cleft constructions are also rare (only two clear cases found).

Prosody

Closed Word Classes

- Explicitly given items are most often deaccentuated.

Open Word Classes

- Given items are not frequently deaccentuated.
- Implictily given items are never focused.
- New items are stressed/focused equally often.

Conclusions and Future Research

Conclusions / Comments

- All observations are preliminary.
- Only one subject fully labelled by more than one labeller. The lack of reliable consensus analyses makes far-reaching conclusions premature.
- Discourse structure is flat, making it more amenable to studies/modelling of topic shifts rather than hierarchical structures.
- Possible explanations as to why given items are not frequently deaccentuated (for open word classes):
  (1) A lot of time passes in-between turns (up to one minute)
  (2) The tendency to deaccentuate given items is smaller when addressing a machine.
- The reason why new items (open word classes) are as often focused as they are stressed might be that they in most cases are proper (city) names, and thus contrasted with other, previously mentioned, proper (city) names.

Future Reasearch

- Data collection of human--human dialogues. Real travel agents in real environment to yield as realistic data as possible. (December 1997.)
- Bionic human--machine data collection. Real speech recognition / speech synthesis to yield realistic speech application data. (January 1998.)
- The data above will permit comparisons between WOZ dialogues, authentic man--machine dialogues, and realistic man--man dialogues.
- Need to test the predictive power of opening and closing discourse markers, and combinations thereof.
- Detailed studies of pitch range, disfluencies and other phenomena not yet studied.
- The analysis will be tuned to other projects at Telia Research, such as automatic dialect recognition, automatic detection of prosodic prominence etc.
- Labelling symbol toolbox needs further evaluation and refinement. Need to include more labellers.

References

Beckman M. & Ayers, G. *Guidelines for ToBI Labelling*, version 3, March 1997. The Ohio State University.

Bruce, G. 1994. Prosodisk strukturering i dialog. *Svenskans beskrivning 20*, Umeå.

Bruce, G., Granström, B., Filipsson, M., Gustafsson, K., Horne, M., House, D., Lastow, B., Touati, P. 1995. Speech Synthesis in Spoken Dialogue Research. *Proceedings of Eurospeech 1995*, vol. 2, pp. 1169--1172.

Bruce. G, Granström, B., Gustafsson, K., House, D., Touati, P. 1994. Modelling Swedish Prosody in a Dialogue Framework. *Proceedings of ICSLP 1994*, pp. 1099--1102, Yokohama, Japan.

Bruce., G. & Granström, B. 1993. Prosodic Modelling in Swedish Speech Synthesis. *Speech Communication* 13, pp. 63--73.

Bruce, G. & Touati, P. 1992. On the analysis of prosody in spontaneous speech with exemplification from Swedish and French. *Speech Communcation 11*, pp. 453--457.

Bruce, G. & Touati, P. 1990. Analysis and Synthesis of Dialogue Prosody. *Proceedings of ICSLP 1990*, vol. 2, pp. 489--492, Kobe, Japan.

Bruce, G., Granström, B., House, D. Prosodic Phrasing in Swedish Speech Synthesis. *Proceedings of the ESCA Workshop on Speech Synthesis*, pp. 125--128, Autrans, France.

MacDermid, C. & Eklund, C. 1997. *Report on the first WOZ Simulation for the SLT-DB project*. Internal report, Telia Research 1997-11-11.

Ostendorf, M., Price, P. & Shattuck-Hufnagel, S. 1997. *Evaluating the Use of Prosodic Information an Speech Recognition and Understanding. Final Report*. Boston University, Massachusetts.

Pitrelli, J., Beckman, M., Hirschberg, J. 1994. Evaluation of Prosodic Transcription Labeling Reliability in the ToBI Framework. *Proceedings of ICSLP 1994*, pp. 123--126, Yokohama, Japan.

Shriberg, E.E. 1994. *Preliminaries to a Theory of Speech Disfluencies*. PhD Thesis. University of California at Berkeley.

Silverman, K., Beckman, M., Pitrelli, J., Ostendorf, M., Price, P., Pierrehumbert, J., Hirschberg, J. 1992. TOBI: A Standard for labeling English prosody. *Proceedings of ICSLP 1992*, vol. 2, pp. 867--870. Banff, Canada.