Discourse markers and the segmentation of spontaneous speech - The case of Swedish men 'but/and/so'

Horne, Merle; Hansson, Petra; Bruce, Gösta; Frid, Johan; Filipsson, Marcus

1999

Link to publication

Citation for published version (APA):

General rights
Copyright and moral rights for the publications made accessible in the public portal are retained by the authors and/or other copyright owners and it is a condition of accessing publications that users recognise and abide by the legal requirements associated with these rights.

- Users may download and print one copy of any publication from the public portal for the purpose of private study or research.
- You may not further distribute the material or use it for any profit-making activity or commercial gain.
- You may freely distribute the URL identifying the publication in the public portal.

Take down policy
If you believe that this document breaches copyright please contact us providing details, and we will remove access to the work immediately and investigate your claim.
Discourse markers and the segmentation of spontaneous speech
The case of Swedish men ‘but/and/so’

Merle Horne, Petra Hansson, Gösta Bruce, Johan Frid and Marcus Filipsson

Prosodic and lexical correlates of ‘clause-like’ and ‘paragraph-like’ boundaries associated with the Swedish discourse marker men ‘but/and/so’ are examined. Men-tokens in spontaneous monologues were labelled as to their boundary-status, first using text-only data. The ‘strong’ tokens (labelled identically by all labellers) were subsequently seen to be correlated with clear differences in the prosodic and lexical parameters examined. This tendency was not found for the corresponding ‘weak’ tokens which were subsequently relabelled using both text and speech nor for the data-base as a whole. A test using a neural network trained using strong tokens is seen to be able to correctly categorize 90% of the strong men-tokens as to their associated boundary-type. The results show that discourse markers along with their prosodic and lexical correlates constitute a constellation of important information for understanding how segmentation of speech is produced and understood.

Introduction

Within the area of speech recognition and understanding, one area of current research is centered around the issue of segmentation of spontaneous speech into ‘clause-like’ and ‘paragraph-like’ units. (This information can be useful e.g. in referent-resolution algorithms and in algorithms for recognizing and synthesizing topic boundaries). As is known, prosodic cues constitute an important source of information on discourse segmentation (Grosz & Hirschberg 1992, Ostendorf et al. 1993). In previous studies, we have analysed the role played in particular by ‘right-edge’ prosodic cues such as phrase accents and final lengthening (Horne et al. 1995, Bruce et al. 1993) in signalling boundaries, particularly in read speech (see also de Pijper & Sanderman 1994).

One problem encountered in investigations on speech segmentation is that the same prosodic parameters (e.g. F0-reset, final lengthening, pause duration) that are used to mark the boundary between prosodic phrases (=clause-like
units) are also used to mark the boundary between speech paragraphs (≈written paragraphs). What is most often involved is a relative difference in the expression of the different prosodic parameters. Thus, it can be difficult to know where one should set the limit for e.g. F0-reset in order to be able to distinguish between a prosodic phrase boundary and a speech paragraph. Furthermore, it is not at all clear if one always can (or should) try to draw a strict boundary between these two since it is not clear that speakers themselves do. Perhaps it is more appropriate to regard segmentation as a gradient parameter where speakers, due to variation in the extent of speech planning, also use varying degrees of clarity in boundary signalling which in turn, is interpreted as having varying degrees of meaningfulness by the listener (see Swerts 1997).

**Discourse markers**

One additional type of information that can be used to facilitate the segmentation of speech into discourse units that we have not earlier investigated is discourse markers/cue phrases that together with prosodic cues as well as other lexical/syntactic cues, mark the beginning of different units of discourse structure (Schiffrin 1987, Mosegaard Hansen 1997). These are cues that are specifically related to the left edge of these units. They constitute local lexical information that can be used, together with other types of information, to determine if one has to do with a prosodic phrase boundary or a higher speech paragraph boundary (see Nakajima & Allen 1993). The prosodic characteristics of the discourse marker itself are also important since cue words are often ambiguous in the sense that the same word can be associated with different discourse functions.

**Swedish men ‘but/and/so’**

This study reports on the Swedish cue-word *men* which can correspond to English ‘but/and/so’ in spontaneous speech. *Men* is classified lexically as a conjunction, a classification which reflects its function within sentence grammar to link together two or more clauses. In this function, *men* expresses a local contradiction. Following are some examples from our data-base which consists of Swedish narrations of a fragment of a silent film:

(1) Sentential *men*:
   
   (a) man kunde läsa nåt litet av det där brevet *men* det var väldigt otydligt
       ’you could read a little bit of the letter but it was very unclear’
(b) först tror man att han är död men det var han inte
‘first you think he’s dead but he wasn’t’

In addition to this clausal or sentential (S) function, men has another function in spoken language, i.e. to introduce a new ‘speech paragraph’ containing a new discourse (D) topic or to return to a previous topic. In this function, it corresponds to English and (then) or but (anyway) or so.

(2) Discourse (D) men:

a) så får han hjälp upp då och då försvinner de här … hotelldirektören med kompani ut … och han släpar sig bort till skåpet där kappan hänger … men då kommer det in en liten tant … som … nja hon är väl någon sån här husmor på hotellet …

‘so he gets help up and then they go out … the hotel manager and company … and he drags himself over to the wardrobe where the coat is hanging … and then a little old lady comes in … who … yea she’s one of those matrons at the hotel …’

(b) … å eh den här gamle mannen … jag vet inte om han bor ihop med sin dotter eller sin fru men i alla fall frun hon bakar nån form av kaka eller tårtan …

‘oh uh that old man … I don’t know if he lives together with his daughter or his wife but anyway his wife is baking some kind of cake or tart …’

If it were possible to distinguish between the two different kinds of men illustrated in (1-2), one could use this information in speech processing algorithms to facilitate segmentation into clause-like and paragraph-like units in spontaneous speech.

Previous studies
In recent years, a number of studies have been published on the function of discourse markers in discourse (Schiffrin 1987, Fraser 1990, Mosegaard Hansen 1997, Byron & Heeman 1997). Few studies, however, have concentrated on the prosodic correlates of these words (see, however, Hirschberg & Litman 1993 and Fretheim 1988).

In Hirschberg & Litman’s (1993) study, they examined both textual and prosodic features of now and attempted to find a set of features that could be used to distinguish between now’s S(entential) function (adverbial) and its D(iscourse) function. In its D(iscourse) function, now, like Swedish men, marks
the beginning of a new speech paragraph. Hirschberg and Litman arrived at
the following results after examining 100 cases of now:

(i) Discourse now constituted most often a phrase on its own (41.3% of the
cases). Sentence now hardly ever constituted a phrase by itself.
(ii) Discourse now appeared most often at the beginning of a phrase
(98.4%). Sentential now appeared most often in non-initial position
(86.5%).
(iii) Discourse now was more often deaccentuated than sentential now.
(iv) Discourse now cooccurred with other cue-words, e.g. well now ...

Since Swedish men is lexically a conjunction, and thus always occurs at the
beginning of an utterance, linear position (phrase-initial/non-initial) cannot be
used as a parameter in distinguishing between men’s functions in discourse as
in the case of English now (see (ii) above). However, one can expect that other
prosodic and lexical correlates could be associated with the two different
categories of men.

Current study
In our investigation of Swedish men, we decided to conduct the study
somewhat differently from Hirschberg & Litman 1993. Four of the authors
(MH, GB, PH, MF) first examined the data using only the written
transcriptions in order to see how many cases of men could be classified as
S(entential) or D(iscourse) using only textual information. (In a few of the
transcriptions, there was some indication of the position of pauses, but no
information on pause strength.) Subsequently, auditory and visual acoustic
information was used to see if prosody would aid listeners in classifying the
tokens of men that were not labelled in the same way by all labellers.

Data, speakers and labelling guidelines
As data, we used 21 spontaneous narrations of a fragment of a silent film (The
Last Laugh). All speakers were from southern Sweden and spoke a variety of
southern Swedish. 15 of the speakers were female and 6 were male.

All together the narrations included 157 tokens of men which were labelled
as either S, D or S/D (in cases where the labeller considered it impossible to
decide on either S or D). Labellers were given guidelines and examples (like
those above in (1) and (2) for categorizing men. Following is a translation of
these guidelines:

S-men expresses some kind of local (topic internal) contradiction.
Often it is a referent/topic under discussion that is contrasted with
another referent. It can even be two verbs or two phrases, e.g. two predicates with the same subject that are contrasted. This local contrast should remain if one replaces men with dock ‘however’ or fast ‘though’ (see examples in (1)).

D-men does not express any local contrast but rather introduces an utterance which begins a new topic or takes up a previous topic. In this function, men can often be left out or replaced with och ‘and’ without changing the meaning. Men often cooccurs with i alla fall ‘anyway’ in this function (see examples in (2)).

If both a S(entential) and a D(iscourse) interpretation seem possible, label men as S/D.

Analysis of data labelled from text-alone
In about 50% of the cases, the labellers were in agreement as to the classification of men in the initial labelling from text-only. That is to say, on the basis of textual information alone, we could classify around half of the cases (81). Of these 81 cases, 41 were classified as D-men, 39 as S-men and 1 as S/D. Following Swerts 1997:515, these cases involve ‘strong’ boundary marking where ‘boundary strength is computed as the proportion of subjects agreeing on a given break’.

Prosodic and lexical analysis
Subsequent to the textual analysis of men, we made a prosodic analysis as well as a lexical analysis of the cases where the labellers were in agreement in order to determine which, if any, of the parameters chosen for study constituted reliable distinguishing characteristics of the two types of boundaries associated with men.

In the prosodic analysis, the following parameters were examined:

(a) Preceding pause: one would expect a relatively longer pause before a D-men than a S-men since a D-men marks the beginning of a new speech paragraph and speech paragraphs (often corresponding to written paragraphs) are most often preceded by longer pauses than the beginning of a new prosodic phrase (often corresponding to a clause or sentence) (Fant & Kruckenber 1989, Strangert 1993, Horne et al. 1995).

(b) F0-reset on men: one would expect a relatively larger F0-reset on D-men than on S-men for the same reason as in (a) (see e.g. Brown et al. 1980, Grosz & Hirschberg 1992, Sluijter & Terken 1993, Swerts & Geluykens 1993).

(c) Duration of men. One would expect that D-men would have a greater duration than S-men since the degree of coherence between discourse men
and what follows is less than that between S-men and that which follows. S-men are often perceived as short and reduced, while D-men are perceived as prominent (drawn out and even accented) reflecting the speaker’s planning of a major new discourse topic. Note that this is just the opposite characterization of now’s discourse form in comparison with its sentential form (Hirschberg & Litman 1993). This is obviously due to the lexical category of the words. Since now is an adverb, it is by default nonreduced in its sentential function, whereas men is a conjunction and by default reduced in the same function. In their discourse functions, then, one would expect them to have the opposite level of reduction.

(d) Phrasing. As in the case of now, one would expect that men in its discourse function of signalling a topic shift would constitute a prosodic phrase on its own, i.e. both preceded and followed by pauses reflecting the extra planning time involved at a topic change.

In the lexical analysis, we looked at the word class of items following men. As Hirschberg & Litman showed in the case of now, in its D(iscourse) function, this marker often cooccurs with other discourse markers/cue phrases e.g. well now. Swedish men is no exception. In its more S(entential), topic internal discourse function, on the other hand, men is often followed by a pronoun referring back to an already introduced discourse referent/topic. The only pronoun one would expect to any extent after D-men in Swedish would be nonreferential det ‘it’, ‘there’ which, like it and there in English occurs at the beginning of ‘presentational’ clauses like Det handlar om … ‘It is about …’ or Det var några barn … ‘There were some children …’. Such syntactic constructions occur at the beginning of a new topic unit (for a review of syntactic and referential cues to topic shift/continuity, see articles by Cumming & Ono as well as Tomlin et al. in van Dijk 1997).

Results

Prosodic correlates

Duration. Measurements of the absolute duration of men revealed that D(iscourse)-men was on the average 310 ms long while S(entential)-men was on the average 170 ms, i.e. 55% the duration of D(iscourse) men. This difference was statistically significant (p-level = 0.002).

F0-reset. Measurements of F0 were made at the end of the word preceding men and in the vowel of men. Of the 41 tokens of D-men, 32 (78%) were characterized by a positive F0-reset. Since the phrase-final F0 level preceding
Table 1. *Men*-tokens labelled from text-only.

**Prosodic correlates**

<table>
<thead>
<tr>
<th></th>
<th>(\bar{x})</th>
<th>s.d.</th>
<th>min</th>
<th>max</th>
<th>t</th>
<th>df</th>
</tr>
</thead>
<tbody>
<tr>
<td><em>Duration</em> (ms)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>D (41)</td>
<td>310</td>
<td>240</td>
<td>40</td>
<td>1170</td>
<td>t=3.2</td>
<td>78</td>
</tr>
<tr>
<td>S (39)</td>
<td>170</td>
<td>130</td>
<td>10</td>
<td>600</td>
<td></td>
<td></td>
</tr>
<tr>
<td><em>F0</em>-reset (glottalized) (ST)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>D (17/41)</td>
<td>13.8</td>
<td>6.0</td>
<td>0.7</td>
<td>25.6</td>
<td>t=2.41</td>
<td>24</td>
</tr>
<tr>
<td>S (9/39)</td>
<td>7.6</td>
<td>6.7</td>
<td>1.8</td>
<td>19.1</td>
<td></td>
<td></td>
</tr>
<tr>
<td><em>F0</em>-reset (non-glottalized) (ST)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>D (15/41)</td>
<td>5.7</td>
<td>5.6</td>
<td>0.1</td>
<td>16.4</td>
<td>t=2.093</td>
<td>30</td>
</tr>
<tr>
<td>S (17/39)</td>
<td>2.2</td>
<td>3.6</td>
<td>0.01</td>
<td>12.9</td>
<td></td>
<td></td>
</tr>
<tr>
<td><em>Preceding pause</em> (ms)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>D (27/41)</td>
<td>750</td>
<td>670</td>
<td>30</td>
<td>2320</td>
<td>t=1.557</td>
<td>40</td>
</tr>
<tr>
<td>S (15/39)</td>
<td>460</td>
<td>350</td>
<td>40</td>
<td>1080</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Men as separate prosodic phrase (pauses before and after)

|                      |     |      |     |     |       |     |
| D 14/41 (34%)        |     |      |     |     |       |     |
| S 0/39 (0%)          |     |      |     |     |       |     |

**Lexical correlates**

<table>
<thead>
<tr>
<th>Following discourse marker (one of following 5 labels)</th>
<th>Following pronoun</th>
</tr>
</thead>
<tbody>
<tr>
<td>D 26/41 (63%)</td>
<td>D 5/41 (12%)</td>
</tr>
<tr>
<td>S 2/39 (5%)</td>
<td>S 24/39 (62%)</td>
</tr>
</tbody>
</table>

*Men* was associated with glottalization in a number of cases, and since glottalization has the effect of lowering F0 (Dilley et al. 1996), we factored out the cases of reset associated with glottalization into a separate group. It is seen in table 1 that the D-*men* tokens without reset are characterized by a mean reset of 5.7 ST which is in line with the size of the reset one would expect at a speech paragraph boundary. The mean reset associated with glottalization is, on the other hand, 13.8 ST. As for the S-*men* tokens, the tokens not associated with glottalization showed a mean reset of 2.2 ST (normal at a speech paragraph internal phrase boundary), whereas those occurring after phrase-final glottalization exhibited a correspondingly larger reset of 7.6 ST. These differences were significant (p-level 0.0238 for glottalized reset and 0.0448 for non-glottalized reset).

**Preceding pause.** Measurements of pause duration revealed that 66% of D(iscourse) *men* (n=27) were associated with a preceding pause that was on the average 750 ms long. Of the S(entential) *men* tokens, 38% (n=15) were
preceded by a pause which was on the average 460 ms long. (It is to be noted that in the measurements of duration, no consideration was taken to rate of speech.) Although the mean difference is quite large, it is not significant (p-level=0.1274).

Prosodic phrasing. As in the case of English now, Swedish D(iscourse) men constituted a separate phrase in 34% of the D-tokens. None of the S-men, on the other hand were characterized in this way.

Lexical correlates
In its discourse function, it was seen, as expected, that men cooccurs with other similar words (conjunctions, pause fillers, particles). Unlike now, however, the discourse markers cooccurring with men come in a position following men instead of preceding it. In 26 of the 41 cases of men labelled as D(iscourse) (63%), one of the 5 following labels was associated with another discourse marker/pause marker: alltså, sedan (så), i alla fall, i varje fall, men, (eh)(just) då(så), eh, ja, så, då, när. Only 2 of the tokens of men labelled as S were so characterized and both of them involved a pause marker (eh). On the other hand, 24 of the 39 tokens of men (62%) labelled as S were followed immediately by a pronoun, while only 5 (12%) of the D-men were so characterized. Of the five, one was the nonreferential pronoun det ‘it’ = Eng. ‘there’ (det var några sådana fattigbarn ‘there were some poor children’). The other four were personal pronouns followed in turn by discourse markers, e.g. han väntade i alla fall på hotellet ‘he waited anyway at the hotel’. Thus, one can hypothesize that the lexical cue for identifying D-men (i.e. a following discourse marker) is quite strong.

Summary of findings for data labelled from text-only
Results of the acoustic analysis of prosodic parameters associated with the D-men and S-men (where measurements from all speakers are pooled together) show significant mean differences in F0-reset and absolute word duration. A clear difference in prosodic phrasing wherein D-men, but not S-men constitute an independent prosodic phrase was also observed. The difference in preceding pause duration for all speakers pooled, however, was not significant, a result in line with that reported in Swerts & Geluykens 1993 for a single speaker. The two categories of men were further strongly associated with different local lexical correlates, i.e. D-men were in over 60% of the cases followed by another discourse marker, while S-men were in over 60% of the
cases followed by a pronoun. In summary, although the labelling of the data in this study was done on text-alone, it is seen that the ‘strong’ boundaries are associated with a whole constellation of cues that serve to mark their function in discourse.

Analysis of data labelled from text-with-speech

On the basis of the results obtained from the acoustic analysis of the tokens of S and D-men labelled from text-alone, we initially expected that by listening to the data and observing the acoustic signal associated with the tokens of men that did not receive a unique label, prosodic information might help labelers in coming to agreement as to the labelling of men. Three of the labellers (MF was not available at this time) thus reexamined all the cases of men (n=75) which had not received a unique label during the first part of the study. 6 of these were discarded since they constituted parts of speech disfluencies. The remaining 69 were examined and each case was individually discussed until labelers agreed on a unique label for each token.

30 tokens were classified as S, 29 as D and 10 as S/D. During the process, the general impression labellers had was that, although access to the speech signal did help in some cases, there was nevertheless a rather high degree of uncertainty as to the classification of men. This is partly reflected in the fact that 10 tokens were still assigned a S/D label, i.e. it was unclear as to how they should be classified even after listening to the speech.

Results

Prosodic correlates

This general uncertainty as to the classification of the cases of men that were unclear from their text-only labelling is reflected in the prosodic analysis of these tokens after reexamination and retagging using speech as well. Results are presented in table 2. As can be seen, none of the prosodic parameters examined exhibited significant differences between S and D-men tokens.

Duration. As can be seen from table 2, there was hardly any difference between S-men and D-men as regards their absolute duration in the labelled-from-speech data (210 vs 222 ms, respectively).
Table 2. Men-tokens labelled from text and speech.

Prosodic correlates

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Duration (ms)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>210</td>
<td>222</td>
<td>10.7</td>
<td>10.4</td>
<td>3.1</td>
<td>3.2</td>
<td>740</td>
<td>540</td>
<td>D 3/29 (10%)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>160</td>
<td>200</td>
<td>6.6</td>
<td>4.1</td>
<td>1.5</td>
<td>4.6</td>
<td>530</td>
<td>420</td>
<td>S 0/30 (0%)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>20</td>
<td>40</td>
<td>0.4</td>
<td>2.6</td>
<td>0.8</td>
<td>0.1</td>
<td>70</td>
<td>60</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>690</td>
<td>920</td>
<td>22.9</td>
<td>14.0</td>
<td>6.8</td>
<td>16.2</td>
<td>1930</td>
<td>1080</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>t=0.204 (57 df)</td>
<td></td>
<td>t=0.119 (16 df)</td>
<td></td>
<td>t=0.812 (21 df)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>p=0.8391</td>
<td></td>
<td>p=0.9064</td>
<td></td>
<td>p=0.4261</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>F0-reset (glottalized) (ST)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>10.7</td>
<td>10.4</td>
<td>0.4</td>
<td>2.6</td>
<td>0.8</td>
<td>0.1</td>
<td>70</td>
<td>60</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>6.6</td>
<td>4.1</td>
<td>22.9</td>
<td>14.0</td>
<td>6.8</td>
<td>16.2</td>
<td>1930</td>
<td>1080</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>p=0.119 (16 df)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>F0-reset (non-glottalized) (ST)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>3.1</td>
<td>3.2</td>
<td>0.8</td>
<td>0.1</td>
<td>6.8</td>
<td>16.2</td>
<td>1930</td>
<td>1080</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>1.5</td>
<td>4.6</td>
<td>6.8</td>
<td>16.2</td>
<td>t=0.587 (23 df)</td>
<td></td>
<td>t=0.812 (21 df)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>0.8</td>
<td>0.1</td>
<td>p=0.9536</td>
<td></td>
<td>p=0.9536</td>
<td></td>
<td>p=0.4261</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Preceding pause (ms)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>740</td>
<td>540</td>
<td>40</td>
<td>60</td>
<td>70</td>
<td>60</td>
<td>1930</td>
<td>1080</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>530</td>
<td>420</td>
<td>1930</td>
<td>1080</td>
<td>t=0.812 (21 df)</td>
<td></td>
<td>t=0.812 (21 df)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>70</td>
<td>60</td>
<td>p=0.4261</td>
<td></td>
<td>p=0.4261</td>
<td></td>
<td>p=0.4261</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Lexical correlates</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Following discourse marker</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>D 9/29 (31%)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>S 5/30 (17%)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Following pronoun</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>D 12/29 (41%)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>S 14/30 (47%)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

F0-reset. Differences in positive F0-reset between S and D-men in the data labelled from listening were not at all significant. After factoring out glottalization, there is almost no difference in the values for F0-reset.

Preceding pause. Differences in preceding pause length for the tokens of men tagged by listening are quite similar to those tagged from text. 17 of the 29 cases labelled as D (59%) were preceded by a pause having a mean duration of 740 ms whereas only 6 of the tokens labelled as S (20%) were associated with a preceding pause which had a mean duration of 540 ms. Thus the pause duration preceding D-men labelled from speech is almost the same as that associated with the D-men labelled from text whereas the mean duration of S-men is 90 ms greater here than for the tokens labelled from text. Thus the overall difference is not as large.

Phrasing. As in the data labelled from text-alone, a clear cue to the classification of men is its status as an independent prosodic phrase
(surrounded by pauses). In the data labelled from text and speech, 3 of the D-men constituted independent prosodic phrases while none of the S-men were so characterized.

**Lexical correlates**
Even the local lexical information associated with the tokens of men labelled from speech was not significantly different in the two cases. Whereas 9 cases (31%) of D-men were followed by another discourse marker, 6 tokens of S-men (17%) were also followed by a discourse marker. Even following pronouns did not help to distinguish the two cases of men. Although 60% of the S-men were followed by a pronoun, 41% of the D-men were also followed by a pronoun.

**Summary**
No clear differences in the values for the lexical and prosodic parameters were seen in this subset of data labelled using text-with-speech. This indicates that these ‘weak’ boundaries are weak even with regard to prosodic parameters and further corroborates the idea developed in Swerts 1997 that, with respect to hearer/reader, not all boundaries are equally meaningful and this is reflected in the degree of ‘strength’ with which they are linguistically realized.

**Analysis of complete database**
A final analysis of the data was made in order to see if any of the prosodic and lexical parameters could be used to distinguish between S- and D-men when pooling both sets of data (labelled from text-alone and labelled from text-with-speech).

Results are presented in table 3. As can be seen, the only parameter which alone shows a significant difference between S- and D-men is men’s absolute duration (p-level = 0.0203). Another clear indication of men’s D-status is its occurrence as a separate prosodic phrase. Although only 24% (n=17) of the D-men constitute a separate prosodic phrase, none of the S-men are characterized in this way.

As regards the lexical correlates, one can see that half of the D-men are followed by another discourse marker and 55% of the S-men are followed by a pronoun. Only 12% of the S-men were followed by another discourse marker. 25% of the D-men, however, were also followed by a pronoun.
Table 3. Men-tokens (complete database).

**Prosodic correlates**

<table>
<thead>
<tr>
<th>Duration (ms)</th>
<th>ĸ</th>
<th>s.d.</th>
<th>min</th>
<th>max</th>
<th>t</th>
<th>df</th>
</tr>
</thead>
<tbody>
<tr>
<td>D (70)</td>
<td>271</td>
<td>217</td>
<td>20</td>
<td>1170</td>
<td>2.347</td>
<td>137</td>
</tr>
<tr>
<td>S (69)</td>
<td>194</td>
<td>169</td>
<td>10</td>
<td>920</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

*F0-reset (glottalized) (ST)*

| D (28/70)          | 12.6 | 6.3  | 0.4 | 25.6| 1.974 | 42 |
| S (16/69)          | 8.8  | 5.8  | 0.2 | 19.1| 0.0549|    |

*F0-reset (non-glottalized) (ST)*

| D (28/70)          | 4.5  | 4.4  | 0.1 | 16.4| 1.796 | 56 |
| S (29/69)          | 2.5  | 3.9  | 0.01| 16.2| 0.0778|    |

*Preceding pause (ms)*

| D (44/70)          | 740  | 610  | 30  | 2320| 1.803 | 63 |
| S (21/69)          | 480  | 360  | 40  | 1080| 0.0761|    |

Men as separate prosodic phrase (pauses before and after)

| D 17/70 (24%)      | S 0/69 (0%) |

**Lexical correlates**

<table>
<thead>
<tr>
<th>Following discourse marker</th>
<th>Following pronoun</th>
</tr>
</thead>
<tbody>
<tr>
<td>(one of following 5 labels)</td>
<td></td>
</tr>
<tr>
<td>D 35/70 (50%)</td>
<td>D 17/70 (25%)</td>
</tr>
<tr>
<td>S 8/69 (12%)</td>
<td>S 38/69 (55%)</td>
</tr>
</tbody>
</table>

Implications for speech recognition/understanding

From a recognition point of view, one could imagine, all other things being equal, that it is necessary to develop some way of clearly classifying all instances of *men* as either S or D. However, it is not clear that it is important or necessary to be able to distinguish between two different kinds of *men*. From a speech understanding point of view, it stands to reason that it is the clearly marked cases (correlated with ‘strong’ boundaries (Swerts 1997, 1998) that should be important for discourse processing since it is these that the speaker probably intends the listener to pay particular attention to. In our study, these are the cases in the labelling from text-alone part where all labellers agreed as to the labelling of *men*. The ‘strong’ D-boundaries are the ones clearly associated with a topic-boundary, while the ‘strong’ S-boundaries are those that are clearly topic-internal. The unclear cases (related to ‘weak’ boundaries), on the other hand, which are not distinctly marked in any way, do not have as clear a function in discourse either and can for the most part probably be
disregarded in speech processing since they are not interpretable/meaningful for the listener.

Following this line of reasoning, we thought it would be interesting to see to what extent, using the prosodic and lexical parameters measured in the study, a neural network could be trained to recognize these clear ‘strong’ cases in the whole database.

**Classification by neural networks**

Neural networks are computer models based on the operation of components of the human brain. The particular strength of neural networks lies in their power to generalise, classify and find patterns in multi-dimensional data. When a neural network is supplied with some measured parameters the task of the network is to map these input features onto a classification state, that is, given the acoustic features mentioned above as input, the neural network can be trained to decide which type of class category (D-men or S-men) they match most closely. Neural networks have been applied successfully to many aspects of speech processing, see e.g. Kohonen 1988 for an approach to speech recognition, Sejnowski & Rosenberg 1986 for speech synthesis, and Johansson 1995 for language acquisition. See Beale & Jackson 1990 for a more general introduction to neural networks.

**Data**

In order to build a classifier with a neural network, the network needs (minimally) two data sets: a *training* set and a *validation* set. In the data described above, there were 80 cases of ‘strong’ boundaries labelled from text only. Three of these were discarded since the F0 reset parameter could not be measured due to glottalisation. The remaining 77 cases were divided into two groups, one consisting of 33 cases and the other of 44 cases. These groups were to be used as training and validation data. Care was taken so as to get an even distribution of speakers in both groups and to include all speakers in both groups as well as having an approximately even distribution of S-men and D-men cases in each group (55% and 48% D-men, respectively).

**Network details**

Two neural networks were designed. One used three input nodes for parameters ‘preceding pause duration’ (ppd), ‘word duration’ (dur) and ‘F0 reset’ (fzr). This net thus only used the prosodic correlates. The other net had two additional input nodes for the lexical correlates ‘following cue word’ (c-w)
and ‘following pronoun’ (pro), thus having a total of five input nodes. Both nets had a hidden layer with two nodes, and an output layer with one node for the classification as D-men or S-men. All neural network processing and simulation was performed with the SNNS (1996) package from IPVR in Stuttgart. The training was performed using the Resilient back propagation (Rprop) scheme.

Since the number of cases is quite small, there is a risk that the actual distribution of data in the two sets affects the outcome of the network’s performance. Therefore the network was trained in two different sessions. In the first session, the smaller data set was used as training data and in the second the larger data set was used as training data. Note, however, that the two training sessions were run independent of each other. In neither case was the validation data included in the training set. The results presented constitute an average of the two different sessions.

Results
When using the three acoustic parameters (ppd, dur, fzr), the rate of correct classifications is 90%. When the lexical features (c-w, pro) are included, the rate of correct classifications decreases to 82%. This might seem counterintuitive: adding more information should increase the network’s performance. The explanation for this result, however, lies no doubt in the fact that the two lexical features only are good indicators of D-men or S-men when they OCCUR. In the cases when they are not close to the target word men the distribution between D-men and S-men is almost 40-60, which introduces uncertainty in the network. Since there are more cases where they do not occur than when they do occur, the number of correctly classified cases decreases. A possible remedy for this would be to use these features only when they occur close to a men-token. However, no such modification has yet been tested, but is a possible future experiment.

Discussion
The combination of the prosodic parameters preceding pause duration, word duration, and F0 reset can predict the status of the cue word men as being D-men or S-men correctly in 90% of the ‘strong’ cases. The tendencies observed in the data can thus be utilized to produce a rather accurate classifier. It also indicates that listeners use an aggregate, rather than one particular feature, when they distinguish between the two categories.
Conclusion

From this study on the function of the discourse marker _men_ ‘and/but/so’ in Swedish and its associated prosodic and discourse correlates, it has become evident that it constitutes an important source of information for marking boundaries in spontaneous speech. In combination with its prosodic and lexical correlates, it can be used to distinguish between two different kinds of boundary, smaller ‘clause-like’ and larger ‘paragraph-like’ units. Results from an exploratory study using a neural network show that it is possible to attain a high degree of recognition of ‘strong’ boundaries by using the discourse marker and the associated parameters chosen for this study. The results are also interesting for speech synthesis since in order to generate cohesive discourse, it is important to be able to model the different kinds of boundaries that occur in natural speech. The present study shows that a whole constellation of prosodic and lexical cues need to be taken into consideration in order to understand how speakers perceive and produce boundaries in spontaneous speech.

Acknowledgements

This research has been supported by a grant from the Swedish HSFR/NUTEK Language Technology Programme and by Telia Research. We are also very grateful to Mechtild Tronnier for assistance with statistical analyses and to Stéphane Di Cesaré for help with programming.

References


