The London–Lund corpus of spoken English

## Description and research

Svartvik, Jan

1990

# The London-Lund Corpus of Spoken English

## Description and Research

edited by
Jan Svartvik

LUND STUDIES IN ENGLISH 82
Editors: Sven Bäckman and Jan Svartvik

Lund
University
Press

LUND STUDIES IN ENGLISH 82
Editors: Sven Bäckman and Jan Svartvik

# The London-Lund Corpus of Spoken English

## Description and Research

edited by
Jan Svartvik

Lund University Press

LUND STUDIES IN ENGLISH

Editors: Sven Bäckman and Jan Svartvik
Department of English
Helgonabacken 14
S-223 62 Lund, Sweden

Publishers: Lund University Press
Box 141
S-221 00 Lund, Sweden



Art nr 20182
ISSN 0076-1451
ISBN 91-7966-126-2 Lund University Press
ISBN 0-86238-256-4 Chartwell-Bratt Ltd

Printed in Sweden
Studentlitteratur
Lund 1990

# Contents

# Preface

The appearance of this book marks the end of two projects: the London-Lund Corpus of Spoken English (LLC) and Text Segmentation for Speech (TESS).

As the subtitle indicates, the projects are documented in two parts, Description and Research. Part I is a description of LLC and includes two appendices, one with information about all the 100 texts in the corpus, the other with a list of publications using material from the Survey of English Usage. It is gratifying to see that LLC, which has become a widely used computerized research tool since it began to be distributed ten years ago, now appears in its *complete* form. This has only been made possible through close association over the years between the three parties involved in London, Bergen and Lund. I want to thank all those colleagues at home and abroad who have played instrumental roles in this effort: first, Randolph Quirk and Sidney Greenbaum in their capacities as Directors of the Survey of English Usage at University College London, which after all is the origin and still remains the home base of the London-Lund Corpus; and next, Knut Hofland and his colleagues in the Norwegian Computing Centre for the Humanities in Bergen who have undertaken, in their usual friendly and efficient manner, the worldwide distribution of LLC under the auspices of ICAME.

Part II reports on some of the research carried out within the project Text Segmentation for Speech (TESS) at the Survey of Spoken English. I want to acknowledge with gratitude the dedicated research work of my Lund colleagues, in particular Bengt Altenberg, Mats Eeg-Olofsson, and Anna-Brita Stenström (now at the University of Stockholm).

The publication of this book, prosodic characters and all, proved to be technically problematic, and I am indebted to Jean Hudson for speedily and cheerfully taking charge of the complicated computer transfer operation to produce camera-ready copy and to Ami Gayle for efficiently helping out with a host of other DTP chores.

Finally, I want to record my debt to our sponsors, without whose support there would never have been a computerized London-Lund Corpus or a TESS project, in particular our chief mentor, the Bank of Sweden Tercentenary Foundation. Crucial financial assistance has also been received from L.M.

Ericsson, the Knut and Alice Wallenberg Foundation, and the Erik Philip-Sörensen Foundation. The printing of this volume has been financed by the Swedish Council for Research in the Humanities and Social Sciences.

Lund, March 1990
*Jan Svartvik*

# Prosodic symbols

The following symbols are used in the prosodic transcription of examples in the text. The symbols are equivalent to those in Svartvik & Quirk 1980, where further details are given.

| TYPE | EXAMPLE | EXPLANATION |
|---|---|---|
| TEXTUAL COMMENTS | [laugh] | Contextual comment |
| | «yes» | Subaudible words |
| | *yes* | Simultaneous talk |
| | +yes+ | Simultaneous talk |
| TONE UNIT | ■ | End of tone unit |
| | ‖ | Onset |
| | {yes} | Subordinate tone unit |
| NUCLEUS | Y$\overset{\searrow}{\text{E}}$S | Fall |
| | Y$\overset{\nearrow}{\text{E}}$S | Rise |
| | Y$\overset{\rightarrow}{\text{E}}$S | Level |
| | Y$\overset{\vee}{\text{E}}$S | (Rise-)fall-rise |
| | Y$\overset{\wedge}{\text{E}}$S | (Fall-)rise-fall |
| | Y$\overset{\searrow}{\text{E}}$S Y$\overset{\nearrow}{\text{E}}$S | Fall-plus-rise |
| | Y$\overset{\nearrow}{\text{E}}$S Y$\overset{\searrow}{\text{E}}$S | Rise-plus-fall |
| BOOSTER | ▻yes | Continuance |
| | ▵yes | Higher than preceding syllable |
| | △yes | Higher than preceding pitch-prominent syllable |
| | Δyes | Very high |
| STRESS | ˈyes | Normal |
| | ˈˈyes | Heavy |
| PAUSE | yes•yes | Brief pause (of one light syllable) |
| | yes-yes | Unit pause (of one stress unit or 'foot') |

# PART I: Description

# 1

# The London-Lund Corpus of Spoken English

*Sidney Greenbaum & Jan Svartvik*

As the name implies, the London-Lund Corpus of Spoken English (LLC) derives from two projects. The first is the Survey of English Usage (SEU) at University College London, launched in 1959 by Randolph Quirk, who was succeeded as Director in 1983 by Sidney Greenbaum. The second project is the Survey of Spoken English (SSE), which was started by Jan Svartvik at Lund University in 1975 as a sister project of the London Survey.

The goal of the Survey of English Usage is to provide the resources for accurate descriptions of the grammar of adult educated speakers of English. For that purpose the major activity of the Survey has been the assembly and analysis of a corpus comprising samples of different types of spoken and written British English. The original target for the corpus of one million words has now been reached, and the corpus is therefore complete.

The Survey has also engaged in devising and conducting elicitation experiments that are primarily intended to supplement data from the corpus. These experiments have focused on features in divided or rare use or whose grammatical status is in question. Such research has been particularly valuable in producing evidence for variation in usage and judgment among native speakers of English. This field of Survey activity, however, will not concern us here (see further Greenbaum 1988: 83-93).

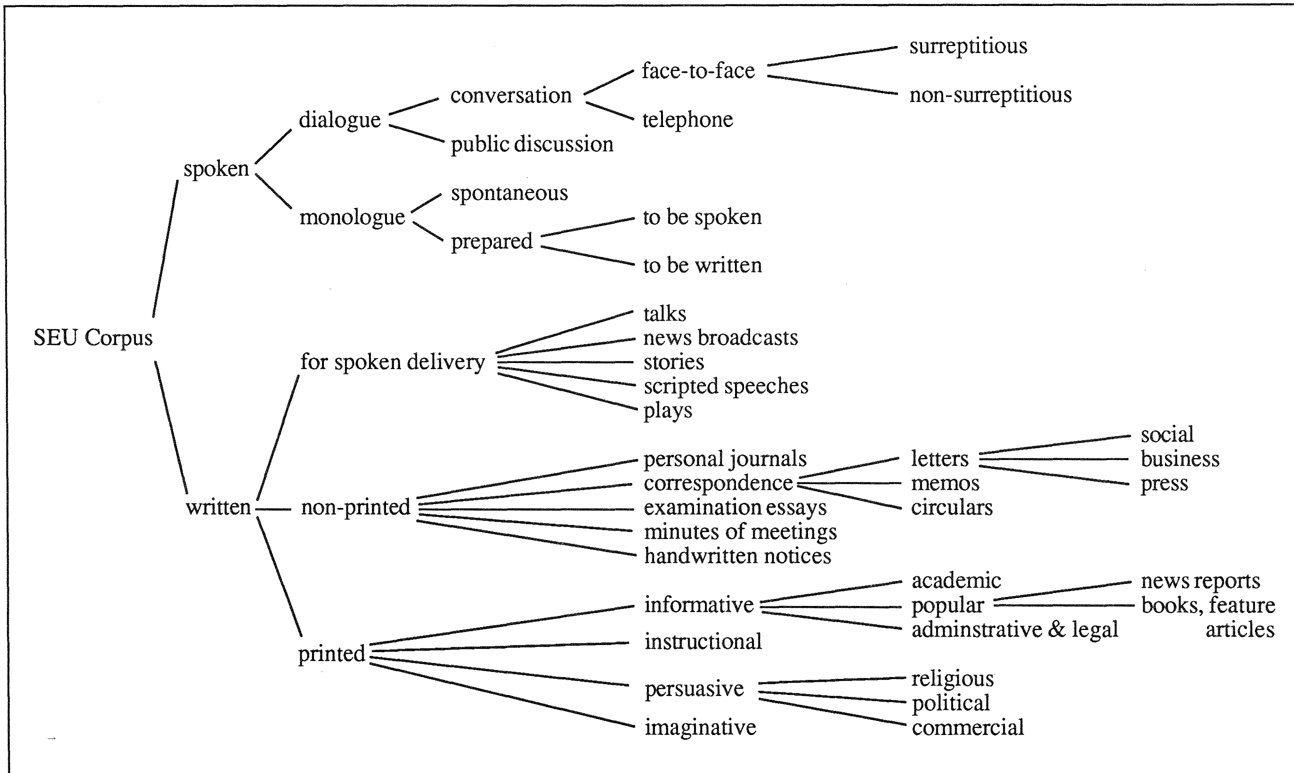
The SEU corpus contains 200 samples or 'texts', each consisting of 5000 words, for a total of one million words. The texts were collected over the last 30 years, half taken from spoken English and half from written English. The spoken English texts comprise both dialogue and monologue. The written English texts include not only printed and manuscript material but also examples of English read aloud, as in broadcast news and scripted speeches. The range of varieties assembled in the whole corpus is displayed in Figure 1:1.

In this book we are interested in the spoken half of the corpus. As can be seen in the figure, the major distinction is between **dialogue** and **monologue**. Within dialogue we distinguish **conversation** in private from **public discussion**. The most common type of conversation is **face-to-face**, which occurs when the participants can see each other and can observe each other's reactions. Technology allows for private conversation by **telephone** when the participants are not in the same place. 'Public discussion' is dialogue that is heard by an audience that does not participate in the dialogue; it includes interviews and panel discussions that have been broadcast. All the telephone conversations and many of the face-to-face conversations were recorded surreptitiously, which means that (at the time of recording) one or more of the participants did not know that their conversation was being preserved. These surreptitiously recorded conversations represent spoken English at its most natural. All the surreptitiously recorded face-to-face conversations with one exception (S.3.7, recorded in 1984) have been published in Svartvik & Quirk (1980).

Within monologue we distinguish **spontaneous** from **prepared**. Spontaneous monologue, which is nearest to conversation in being relatively unplanned, includes running commentaries on sport events and state occasions, demonstrations of experiments, and speeches in parliamentary debates. Prepared monologue, on the other hand, is closest to written English but retains some spontaneity in not being read from a script and therefore allowing for improvisation. Typical prepared monologues in the corpus are sermons, lectures, addresses by lawyers and a judge in court, and political speeches. A special type of prepared monologue is represented by the text of dictated letters, where the speech is intended to be written down.

The spoken corpus of the Survey of English Usage has been transcribed with a sophisticated marking of prosodic and paralinguistic features. All the SEU texts, written as well as spoken, have been analysed grammatically. The grammatical analysis and the prosodic/paralinguistic analysis are represented in the Survey files by typed slips (6x4 inches). Each slip contains 17 lines, including 4 lines of overlap between that slip and the adjacent ones before and after. For each grammatical, prosodic and paralinguistic feature there is one slip that is marked for that item. The Survey collects 65 grammatical features,

**Figure 1:1**. Corpus of the Survey of English Usage.



over 400 specified words or phrases, and about 100 prosodic and paralinguistic features.

In 1975 the Survey of Spoken English was established at Lund. Its initial aim was to make available, in machine-readable form, the spoken material which by then had been collected and transcribed in London: 87 texts totalling some 435 000 words (see Svartvik et al 1982 for an account of the input procedures). The material was inserted in a reduced transcription and without grammatical analysis. Early in 1980 the first copies of the computerized London-Lund Corpus of Spoken English were distributed to interested scholars all over the world.

This **original** London-Lund Corpus of 87 texts (often referred to as **LLC**) has since been augmented by the remaining 13 spoken texts of the SEU corpus, which were processed at the Survey of English Usage in conformity with the system used in the original London-Lund Corpus. These 13 texts constitute a **supplement** (**LLC:s**) to the original computerized version. The **complete** London-Lund corpus (**LLC:c**) therefore consists of 100 spoken texts. In addition, all the written texts of the SEU corpus are now computerized, but these do not form part of the London-Lund Corpus and will not be distributed, though they can be consulted at the Survey of English Usage at University College London. Since LLC has been widely used in scholarly publications for the last decade, it is important to distinguish in future publications the original version from the supplement and from the complete version that incorporates the supplement. In order to avoid misunderstanding we recommend using suffixes for all three thus:

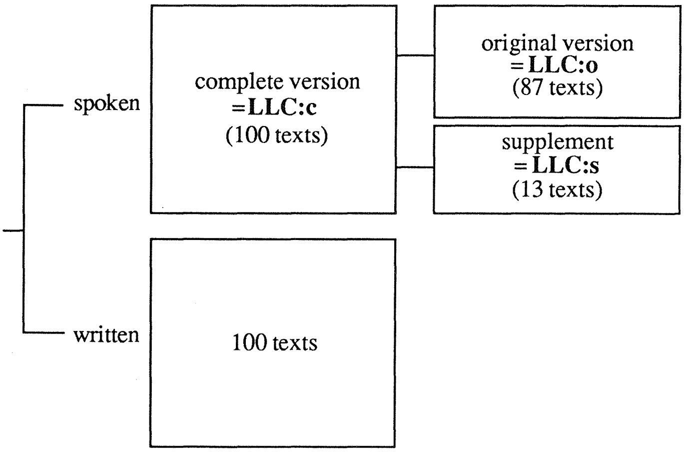
LLC:o the original corpus (87 texts)
LLC:s the supplement (13 texts) to the original corpus
LLC:c the complete corpus (100 texts)

The constituents of the complete SEU corpus are displayed in Figure 1:2. Appendix 1 lists all the 100 spoken texts of LLC:c in order of text category, and provides (as far as the information is available) the dates of recordings and certain bibliographical details about the speakers.

Within the written SEU corpus, 17 texts were recorded from spoken deliveries of written material, such as news broadcasts, plays, and scripted speeches. These are not included in LLC:c, though in the computerized version they have been transcribed in the same way as the spoken texts.

We must distinguish the full prosodic and paralinguistic transcription in the SEU corpus from the reduced transcription in LLC:c and in the computerized 17 texts that were read aloud from written material.

**Figure 1:2.** The computerized SEU corpus.




The basic prosodic features marked in the **full transcription** are tone unit boundaries, the location of the nucleus (ie the peak of greatest prominence in a tone unit), the direction of the nuclear tone, varying lengths of pauses, and varying degrees of stress. Other features comprise varying degrees of loudness and tempo (eg allegro, clipped, drawled), modifications in voice quality (pitch range, rhythmicality and tension), and paralinguistic features such as whisper and creak. Indications are given of overlap in the utterances of speakers. The full transcription and the grammatical analysis are available only on the slips at the Survey of English Usage at University College London.

The **reduced transcription** of the computerized LLC:c corpus and the 17 computerized texts of written English read aloud retains the basic prosodic features of the full transcription but omits all paralinguistic features and certain indications of pitch and stress. It retains the following features: tone units (including the subdivision where necessary into subordinate tone units), onsets (the first prominent syllable in a tone unit), location of nuclei, direction of nuclear tones (falls, rises, levels, fall-rises, etc), boosters (ie relative pitch levels), two degrees of pause (brief and unit pauses alone or in combination) and two degrees of stress (normal and heavy). Also indicated are speaker identity, simultaneous talk, contextual comment ('laughs', 'coughs', 'telephone rings', etc) and incomprehensible words (ie where it is uncertain what is said in the recording). For explanations of the prosodic and paralinguistic system we refer to Crystal 1969. Researchers may obtain from the Survey of English

Usage a guide to the full SEU transcription and an account of differences between the full and the reduced transcriptions. Figure 1:3 presents a sample slip with the full transcription, and Figure 1:4 gives the same passage (tone units 139-163 of text S.1.3) in reduced transcription as printed in Svartvik & Quirk 1980:85.

**Figure 1:3.**

```
                      S.1.3-9
                      *bro*chùre* for# so I q /dı́d it#q# . and /then ¯ anóther
                      *one#p#  - and
                   B   *  mhm *
                  (A) /Nthèn they ˚said# well "/now that you've done Nthése#
                      and they've been a "/sò sucNcéssful#a# we'd /like you
                      to do our l Nsùpēr# . /alpha:màtic#l# a or /Nsómething#
m:narrow              and /this is one of Nthése#a# that /goes mm' ¯sı́deways#
m':rhythmic           and /fróntwards# and ¯em/bróiders# and */dárns#m'# and
                      sews */bùttons on#m#
                   B  *( - laughs ) yes*
                  (A) -- a and I /saìd# well a# I /don't Nrèally ˚think# I
m:slurred             could /write# -- a and this was a m sort of m#a#
                      /ninety six page :bòoklet# p /you Nknów# p# about /that
                      Nbìg# *-* əm I'd I'd /used to gò through# /each of the
                   B  * m *
                  (A)*processes at :hòme# *.* a I ¯don't think it will be
                      *e/nough a# just to have
```

**Figure 1:4.**

| | |
|---|---|
| | BRO☆CHÙRE☆ for■ 139 so I ‖DÍD it■ · 140 and ‖then ANÓTHER one■ – |
| | 141 and |
| b | 142 ☆[mhm]☆ |
| >A | 141 ‖THÈN they ▻said■ 143 well ↑‖now that you've done THÉSE■ 144 and they've been ↑‖SÒ SUCCÉSSFUL■ 145 we'd ‖like you to do our SÙPER■ · 146 ‖ALPHA▵MÀTIC■ 147 or ‖SÓMETHING■ 148 and ‖this is one of THÉSE■ 149 that ‖goes SÍDEWAYS■ 150 and ‖FRÓNTWARDS■ 151 and EM‖BRÓIDERS■ 152 and ☆‖DÁRNS■ 153 and sews☆ ‖BÙTTONS on■ |
| b | 154 ☆( – laughs) yes☆ |
| >A | 155 – – and I ‖SAÌD■ 156 well I ‖don't RÈALLY ▻think■ 157 I could ‖WRÌTE■ – – 158 and this was a sort of ‖ninety-six page ▵BÒOKLET■ 159 ‖you KNÓW■ 160 about ‖that BÌG■ ☆–☆ 161 [əm] I'd I'd ‖need to GÒ through■ 162 ‖each of the |
| b | 163 ☆[m]☆ |
| >A | 162 processes at ▵HÒME■ ☆·☆ 164 I don't think it will be e‖nough just to have |

There is a concordance of LLC:o, and both the text and the concordance are available from the International Computer Archive of Modern English (ICAME). We hope that the complete spoken corpus (LLC:c) will be available from ICAME in the near future.

The address of ICAME, where copies of the London-Lund Corpus can be obtained, is:

The Norwegian Computing Centre for the Humanities
University of Bergen
P.O. Box 53
N-5027 Bergen, Norway

The address of the Secretary of ICAME and the Editor of *ICAME Journal* is:

Professor Stig Johansson
Department of English
University of Oslo
P.O. Box 1003, Blindern
N-0315 Oslo 3, Norway

The spoken texts in the full or reduced version have been extensively used by scholars throughout the world in studies of spoken English and in comparisons between the spoken and written language. For computer studies, comparisons have been drawn between the London-Lund Corpus and two corpora of printed texts dating from 1961, each consisting of about one million running words: the Lancaster-Oslo/Bergen (LOB) British corpus and the Brown University American corpus. Comparative studies generally refer to frequency differences within and across the corpora (see Chapter 2).

The Survey of English Usage holds a list of publications that have used Survey material, including LLC, and updates the list annually. The list is available to scholars on request. Appendix 2 shows that well over 200 publications have used Survey material, and no doubt there are more that are not known to us. They consist of general works (such as grammars), monographs, chapters in books, and articles. Prominent among the books is the standard reference grammar of modern English - *A comprehensive grammar of the English language* (Quirk et al 1985).

APPENDIX 1

# The complete London-Lund Corpus

Below we provide basic information on all the 100 spoken texts (ie the **complete** version of the London-Lund Corpus, above referred to as 'LLC:c'), including text category (eg *conversation*), year of recording (eg *1984*), speaker category (eg *female undergraduate*) and speaker age (eg *c. 20*). In this complete listing, the 87 texts included in the **original** London-Lund Corpus are marked 'LLC:o'; the 13 texts which constitute a **supplement** are marked 'LLC:s'; the 34 texts available in print (Svartvik & Quirk 1980) are marked 'CEC'.

Unless indicated otherwise, the speakers are British. In the recordings made without prior knowledge of the main participants, the names used in the transcriptions are fictitious but prosodically equivalent to the originals. Speakers denoted by upper case letters have been surreptitiously recorded. Sometimes one or more participants had knowledge of the recording (and had the task of keeping the conversation going); such speakers, whose contributions have not been prosodically transcribed, have been specially designated by lower case letters. Some of the texts are composite and contain 'subtexts', either with the same speakers, or recorded in a comparable setting.

**Texts S.1-2: Conversations between equals**

Text S.1.1 (1964) LLC:o+CEC

- **A** male academic, c. 44
- **B** male academic, c. 60

Text S.1.2 LLC:o+CEC

S.1.2 (1963)

- **A** male academic, c. 43
- **B** male academic, c. 42

S.1.2a (1965)

- **A** male academic, c. 45
- **B** male academic, 41
- **CAL** telephone caller

S.1.2b (1965)

- **A** male academic, 45
- **B** male academic, 36

Text S.1.3 (1965) LLC:o+CEC

- **A** female undergraduate, c. 36
- **b** female undergraduate, c. 30
- **c** male undergraduate, c. 36

Text S.1.4 (1969) LLC:o+CEC

- **A** male academic, c. 48
- **B** male academic, c. 48

Text S.1.5 (1967) LLC:o+CEC

- **A** female secretary, c. 21
- **B** female academic, c. 25
- **C** female secretary, c. 35
- **D** female secretary, c. 21

Text S.1.6 (1964) LLC:o+CEC

- **A** female academic, c.45
- **B** male academic, c. 28

Text S.1.7 (1972) LLC:o+CEC

- **a** male academic, mid 30s
- **A** male primary school teacher, c. 30
- **B** male secondary school teacher, c. 30

Text S.1.8 (1969) LLC:o+CEC

**A** female academic, c. 55

**B** female academic, c. 50

**C** female academic, c. 23

Text S.1.9 (1966) LLC:o+CEC

**a** male academic, c. 40

**A** female academic, c. 30

**B** male academic, c. 40

**C** male academic, c. 55

Text S.1.10 (1975) LLC:o+CEC

**A** female lecturer, c. 52

**b** female academic, c. 40

**c** businessman, c. 52

Text S.1.11 (1975) LLC:o+CEC

S.1.11a

**A** housewife, c. 60
(mother of **c**, future mother-in-law of **B**)

**B** male computer specialist, c. 30

**c** female research worker, c. 20

S.1.11b

same speakers as in S.1.11a

Text S.1.12 (1975) LLC:o+CEC

**a** female academic, c. 40

**B** housewife, c. 50

**c** businessman, c. 50

**D** male research chemist, c. 51

Text S.1.13 (1975) LLC:o+CEC

**a** businessman, c. 50

**B** housewife, c. 60 (wife of **C**)

**C** male retired charity commissioner, c. 61

Text S.1.14 (1976) LLC:o+CEC

S.1.14a

**A** male academic, 60

**b** businessman, c. 52

**c** female academic, c. 40

S.1.14b

same speakers as in S.1.14a

Text S.2.1 LLC:o+CEC

S.2.1 (1963)

**a** male academic, c. 43

**B** male academic, c. 34

S.2.1a (1953)

**a** male academic, c. 33

**B** male academic, c. 25

S.2.1b (1953)

same speakers as in S.2.1a

**a** is the same speaker in all three subtexts

Text S.2.2 (1969) LLC:o+CEC

S.2.2a

**a** male academic, c. 48

**B** male stockbroker, 35

S.2.2b

**a** male academic, c. 40

**A** male company employee, c. 40

Text S.2.3 (1974) LLC:o+CEC

**a** female academic, c. 40

**A** male legal civil servant, c. 40

**B** male architect, c. 43

Text S.2.4 (c. 1970) LLC:o+CEC

S.2.4a

**A** male academic, c. 35

**B** male academic, c. 30

**c** housewife, c. 30

**d** male academic, c. 35 (husband of **c**)

S.2.4b

same speakers as in S.2.4a

Text S.2.5 (1974) LLC:o+CEC

S.2.5a

**A** male academic, c. 26

**B** female academic, c. 48

**C** female academic, c. 62

**a** American male academic, c. 40

S.2.5b

same speakers as in S.2.5a

Text S.2.6 (1974) LLC:o+CEC

**a** male academic, c. 55

**A** male academic, c. 50

**B** male academic, c. 50

**C** male academic, 50-55

Text S.2.7 (1975) LLC:o+CEC

**a** male academic, c. 20

**b** female teacher, c. 20 (wife of **a**)

**C** female studio manager, 20

Text S.2.8 (1975) LLC:o+CEC

S.2.8a

**A** male civil engineer, 34

**B** male civil engineer, 38

**f** female academic, 49

**m** businessman, 60

S.2.8b

**a** same speaker as **m** in S.2.8a

**A** same speaker as **A** in S.2.8a

**B** same speaker as **B** in S.2.8a

**C** same speaker as **f** in S.2.8a

Text S.2.9 (1974) LLC:o+CEC

**a** male academic, c. 27

**A** male doctor, c. 29

**B** female secondary school teacher, c. 27

Text S.2.10 (1975) LLC:o+CEC

**A** male merchant banker, c. 30

**B** housewife, c. 30 (wife of **A**)

**c** male computer specialist, c. 30

**C** same speaker as **c** (but from TU 1113 on he is no longer aware of being recorded and is called **C**)

**d** female research worker, c. 25

Text S.2.11 (1975) LLC:o+CEC

S.2.11a

**A** male computer specialist, c. 30

**b** female research worker, c. 25

S.2.11b

**A** female employee, c. 20

**B** male employee, c. 20

**c** same speaker as **A** in S.2.11a

**d** same speaker as **b** in S.2.11a

Text S.2.12 (1975) LLC:o+CEC

**a** female teacher, c. 25

**A** female medical nurse, 23

Text S.2.13 (1976) LLC:o+CEC

**A** male social worker, late 20s

**B** female social worker, late 20s

**c** male computer specialist, 31

**d** female research assistant, 28

Text S.2.14 (1976) LLC:o+CEC

**A** female science graduate, secretary, 52

**b** businessman, c. 50

**C** female research assistant and author, c. 50

**Texts S.3: Conversations between disparates**

Text S.3.1 (1961) LLC:o+CEC

S.3.1a

**a** male academic, c. 40

**A** female prospective undergraduate, c. 20

**B** male academic, c. 40

S.3.1b

**a** same speaker as in S.3.1a

**A** female prospective undergraduate, c. 20

**B** same speaker as in S.3.1a

S.3.1c

**a** same speaker as in S.3.1a

**A** female prospective undergraduate, c. 20

**B** same speaker as in S.3.1a

Text S.3.2 LLC:o+CEC

S.3.2a (1973)

**A** male academic, c. 52

**B** female ex-research assistant, c. 30

S.3.2b (1974)

**A** male academic (former employer), 54

**B** male academic (former employee), c. 28

S.3.2c (1975)

**A** male academic, c. 50

**B** female academic, c. 30

Text S.3.3 (c. 1971) LLC:o+CEC

**A** male administrator, c. 55

**B** male undergraduate, c. 20

**C** female undergraduate, c. 20

**D** male undergraduate, c. 20

**E** female undergraduate, c. 20

**F** male undergraduate, c. 20

Text S.3.4 (c. 1971) LLC:o+CEC

**A** male administrator, c. 55

**B-F** male academics, 45-60

Text S.3.5 (1961) LLC:o+CEC

S.3.5a

**a** male academic, c. 40

**A** male prospective undergraduate, c. 18

**B** male academic, c. 40

S.3.5b

**a** same speaker as in S.3.5a

**A** male prospective undergraduate, c. 18

**B** same speaker as in S.3.5a

Text S.3.6 (1974) LLC:o+CEC

**a** male academic, c. 50

**A** male academic, c. 50

**B** male academic, c. 30

**C** male academic, c. 50

**D** male academic, c. 50

**E** male academic, c. 48

**F** female academic, c. 55

Text S.3.7 (1984) LLC:s

- **A** male architect, c. 35
- **B** male client, 40
- **c** female client, 36

**Texts S.4: Conversations or discussions between equals**

Text S.4.1 (1969) LLC:o

- **a** male undergraduate, 25
- **b** housewife and teacher, 24 (married to **a**)

Text S.4.2 (1971) LLC:o

- **a** male solicitor, c. 40
- **b** female academic, c. 26

Text S.4.3 (1972) LLC:o

- **A** housewife, c. 35
- **B** housewife, c. 30
- **c** male academic, c. 32 (husband of **B**)
- **D** male academic, c. 35 (husband of **A**)

Text S.4.4 (1975) LLC:o

- **A** male retired civil servant, c. 70
- **b** female language teacher, c. 25
- **c** male computer specialist, early 30s (husband of **b**)
- **D** male teacher, 28

Text S.4.5 (1976) LLC:o

- **a** male postgraduate, 25
- **b** female hospital management student, 19
- **c** female research worker, c. 65

Text S.4.6 (1976) LLC:o

S.4.6a

- **a** male postgraduate, late 20s
- **B** female language researcher, 40s
- **c** male research assistant, late 20s
- **d** female secretary, early 20s

S.4.6b

**A** female librarian, mid 20s
**B** female teacher, mid 20s
**C** male computer specialist, mid 20s
**d** female research worker, mid 20s (wife of **C**)
**e** male computer specialist, 30

Text S.4.7 (1976) LLC:o

**a** female actress, 20
**b** female research assistant, c. 65
**c** male graphic artist, 28 (husband of **a**, son of **b**)

### Texts S.5: Radio discussions and conversations between equals

Text S.5.1 (1959): Radio discussion LLC:o

**f** male broadcaster, 71
**h** male novelist, 51
**m** male politician, 51
**jl** male politician, 54
**tl** male educationist, 40

Text S.5.2 (1958): Radio discussion LLC:o

**f** male educationist, 38
**d** male academic, 36
**g** male academic, 46
**k** male academic , 49
**m** male author, 39

Text S.5.3 (1961): Radio discussion LLC:o

**n** male politician, 48
**w** male academic, 42
**l** male journalist, c. 35

Text S.5.4 (1958): Radio discussion LLC:o

**f** male broadcaster, 76
**h** female journalist, c. 43
**d** male author and journalist, 58
**c** male journalist, 38
**v** male lawyer, 49

Text S.5.5 (1960): Radio discussion LLC:o

**a** male broadcaster, 72
**b** male politician, 60
**c** male politician, 29
**e** male politician, 35
**f** male politician, 45
**d, g, j, k** female & **h** male speakers from the audience

Text S.5.6 (1961): Radio discussion LLC:o

**b** male banker and farmer, 55
**m** male author and astronomer, 38
**d** male veterinary surgeon
**c** male compère

Text S.5.7 (1970): Radio discussion LLC:o

**a** male journalist, 45
**b** male head of college, c. 45
**c** female academic, c. 45
**d** male academic, c. 45

Text S.5.8 (1971): Private conversation LLC:o

**a** female postgraduate student, c. 26
**b** female postgraduate student, c. 21

Text S.5.9 (1971): Private conversation LLC:o

**a** male academic, 23
**b** female secretary, 22

Text S.5.10 (1971): Private conversation LLC:o

**a** male English teacher, c. 29
**b** male English teacher, c. 30

Text S.5.11 (1976): Private conversation LLC:o

S.5.11a

**a** male senior academic, 50s
**B** male senior academic, 50s

S.5.11b

same speakers as in S.5.11a

Text S.5.12 (1985): Committee meeting LLC:s

- **a** female teacher, 30-40
- **b** male conductor, 30-40
- **c** female academic, 30-40
- **d** male computer specialist, 30-40
- **e** female academic, 30-40
- **f** female administrator, 30-40
- **g** female TV producer, 30-40
- **h** male civil servant, 30-40

Text S.5.13 (1986): Academic meeting LLC:s

- **a** male academic, 50
- **b** male academic, 60
- **c** unknown
- **d** male academic, 42
- **e** male academic, 45
- **f** male academic, 38
- **g** male academic, 57
- **h** female academic, c. 60
- **j** male academic, 63
- **k** male academic, 54
- **l** male academic, 49
- **m** male academic, 62
- **n** male academic, c. 60
- **p** male academic, 54
- **q** inaudible

## Texts S.6: Interviews and conversations between disparates

Text S.6.1 (1966): Radio interviews LLC:o

S.6.1a

- **a** female broadcaster, c. 25
- **b** female academic, c. 25

S.6.1b

same speakers as in S.6.1a

S.6.1c

- **a** same speaker as **a** in S.6.1a
- **b** male academic, c. 40

Text S.6.2 (1961): Private conversation LLC:o

- **a** male academic, c. 40
- **b** female ex-student, c. 24

Text S.6.3 (1974): Radio interview LLC:o

- **a** male broadcaster, 43
- **b** male politician, c. 60

Text S.6.4 LLC:o

S.6.4a (1975): Radio discussion

- **b** male retired lawyer, father of **a**, 95
- **a** male radio producer, son of **b**, 53
- **c** female, daughter of **a**, 24

S.6.4b (1973): Conversation

- **a** female research worker, 47
- **b** female ex-editor, 85

Text S.6.5 (1975): Radio discussion LLC:o

- **j** female broadcaster, c. 54
- **w** female counsellor, c. 52
- **p** male counsellor, c. 35
- **g** male graduate architect, counsellee, 50
- **m** female counsellee, c. 50 (wife of **g**)

Text S.6.6 (1974): Monologue within a radio interview LLC:o

- **a** female former nurse, 87

Text S.6.7 (1971): Radio interview with an elder statesman LLC:s

- **a** interviewer, 54
- **b** elder statesman, 77

Text S.6.8 (1977): Psychiatrist's discussion group LLC:s

- **a** female senior social worker, c. 40
- **B** female psychotherapist, c. 50 (non-native, not transcribed prosodically)
- **c** male psychiatrist, c. 40
- **d** female social worker, c. 25

Text S.6.9: Computer lesson LLC:s

S.6.9a (1985): Use of computer in library cataloguing

- **a** male librarian, 30-40
- **b** female librarian, 20s

S.6.9b (1987): Use of computer for linguistic processing

**p** male computer expert, c. 50

**q** male academic, c. 65

### Texts S.7: Telephone conversations between equals

Text S.7.1: Telephone conversations LLC:o

S.7.1a (1967)

**A** female telephone speaker

**b** male university lecturer, c. 35

**C** female research assistant, c. 25

S.7.1b (c. 1961)

**B** female training college lecturer, c. 40

**c** male research assistant, c. 35

S.7.1c (c. 1961)

**A** female university secretary

**b** female university secretary, c. 25

S.7.1d (c. 1961)

**a** female university secretary

**B** female university secretary, c. 25

S.7.1e (c. 1966)

**a** male university lecturer, c. 35

**B** female graduate, c. 35

Text S.7.2 (1975): Telephone conversations LLC:o

S.7.2a

**A** female researcher, mid 40s

**B** male postgraduate, late 20s

S.7.2b

**A** female teacher, 20s

**B** male research assistant, late 20s

S.7.2c

**A** female administrator, 29

**B** female university secretary, early 20s

S.7.2d

**A** male researcher, 60s

**B** female, 60 (wife of **A**)

S.7.2e

**B** male university lecturer, 50s

**A** female, c. 50 (wife of **B**)

S.7.2f

**A** male computer specialist, early 30s

**B** female researcher, c. 25 (wife of **A**)

S.7.2g

same speakers as in S.7.2b

S.7.2h

**A** male teacher, 36

**B** female researcher, c. 25

S.7.2i

**A** female university secretary, early 20s

**B** female administrator, 29

S.7.2j

**B** female university lecturer, c. 40

**A** male (friend of **B**)

S.7.2k

**A** male university lecturer, c. 40

**B** female university lecturer, c. 40

S.7.2l

**A** female administrator, 29

**B** female university secretary, early 20s

S.7.2m

**A** female university secretary, early 20s

**B** female administrator, 29

S.7.2n

**A** female administrator, 29

**B** female university secretary, early 20s

S.7.2o

**A** telephonist

**B** female researcher, mid 40s

**C** female language teacher

S.7.2p

**A** female researcher, c. 25

**B** male journalist, 30

Text S.7.3 (1975): Telephone conversations LLC:o

S.7.3a

**A** female university secretary, early 20s

**B** female administrator, 29

S.7.3b

**A** female administrator, 29

**B** female university secretary, early 20s

S.7.3c

**A** female university secretary, early 20s

**B** female administrator, 29

S.7.3d

**A** male architect consultant, 30

**B** female researcher, c. 25

S.7.3e

**B** female university secretary, early 20s

**A** male, c. 28 (friend of **B**)

S.7.3f

**A** female researcher, c. 25

**B** male journalist, 30

S.7.3g

**A** male postgraduate researcher, late 20s

**B** female (wife of **A**)

S.7.3h

**A** female

**B** female university secretary, early 20s (daughter of **A**)

S.7.3i

**A** female

**B** female university secretary, early 20s (daughter of **a**)

S.7.3j

**A** female university secretary, early 20s

**B** female, 22 (sister of **A**)

S.7.3k

**A** male postgraduate researcher, late 20s

**B** male college lecturer, c. 30

S.7.31

**A** female researcher, c. 25

**B** male computer specialist, 32

S.7.3m

**A** female university secretary, early 20s

**B** female administrator, 29

## Texts S.8: Telephone conversations between equals

Text S.8.1 (1975): Telephone conversations between business associates LLC:o

S.8.1a

**A** male broker

**B** female university lecturer, c. 30

S.8.1b

**A** male broker

**B** female university lecturer, c. 30

S.8.1c

**A** female secretary

**B** female university secretary, 20s

**C** female secretary

S.8.1d

**A** male estate agent

**B** female university lecturer, c. 30

S.8.1e

**A** sewing machine salesman

**B** female university researcher, c. 25

S.8.1f

**A** female university secretary, 20s

**B** female senior secretary

S.8.1g

**A** female telephonist

**B** female university secretary, 20s

S.8.1h

**A** female telephonist

**B** female university secretary, 20s

**C** female secretary

S.8.1i

**A** female clerk

**B** female university secretary, 20s

S.8.1j

**A** female university secretary, 20s

**B** female speech therapist

S.8.1k

**A** female secretary

**B** female university secretary, 20s

S.8.1l

**A** female telephonist

**B** female university secretary, 20s

**C** female secretary

S.8.1m

**A** female clerk

**B** female university secretary, 20s

S.8.1n

**A** female telephonist

**B** female university researcher, c. 25

**C** male official

S.8.1o

**A** female administrator

**B** female university secretary, 20s

S.8.1p

**A** female telephonist

**B** female university lecturer, c. 30

**C** male estate agent

Text S.8.2 (1975): Telephone conversations between business associates — LLC:o

S.8.2a

**A** female telephonist

**B** female university lecturer, c. 30

**C** male estate agent

S.8.2b

**A** female clerk

**B** female university secretary, 20s

**C** female clerk

S.8.2c

**A** male enquirer

**B** female university secretary, 20s

S.8.2d

**A** female university secretary, 20s

**B** female telephone enquiry clerk

S.8.2e

**A** female telephonist

**B** female university secretary, 20s

**C** female telephonist

**D** female secretary

S.8.2f

**A** female secretary

**B** female university secretary, 20s

S.8.2g

**A** female telephonist

**B** female university lecturer, c. 30

**C** male estate agent

S.8.2h

**A** female secretary

**B** female university secretary, 20s

Text S.8.3 (1975): Telephone conversations between business associates LLC:o

S.8.3a

**A** female secretary

**B** female university secretary, 20s

S.8.3b

**A** female telephonist

**B** female university lecturer, c. 30

**C** male mortgage consultant

S.8.3c

**A** female secretary

**B** female university secretary, 20s

S.8.3d

**A** female university secretary, 20s

**B** male caller

S.8.3e

**A** female official

**B** female secretary

**C** male university researcher, 60s

S.8.3f

**A** male electrician

**B** male research assistant, late 20s

S.8.3g

**A** female personal assistant

**B** female university researcher, 40s

S.8.3h

**A** male house agent

**B** female secretary

**C** female academic

S.8.3i

**A** female secretary

**B** female university researcher, 40s

S.8.3j

**A** female telephonist

**B** male publisher

**C** male academic, 44

S.8.3k

**A** female university secretary, 20s

**B** female secretary

S.8.3l

**A** same speaker as **A** in S.8.3k

**B** female official

S.8.3m

**A** female secretary

**B** female university secretary, 20s

Text S.8.4 (1975-6): Telephone conversations between business associates LLC:o

S.8.4a

**A** female university researcher, 40s

**B** male research assistant, late 20s

S.8.4b

**A** same speaker as **A** in S.8.4a

**B** same speaker as **B** in S.8.4a

S.8.4c

**A** male academic

**B** male friend of A

S.8.4d

**A** male, 30-40

**B** female university secretary, 20s

S.8.4e

**A** female university secretary, 20s

**B** female university researcher, 40s

S.8.4f

**A** female secretary

**B** female (wife of **B** in S.8.4a)

S.8.4g

**A** female university researcher, 40s

**B** male university researcher, 40s

S.8.4h

**A** same speaker as **A** in S.8.4g

**B** male administrator, 50s

S.8.4i

**A** same speaker as **A** in S.8.4g

**B** female architect

S.8.4j

**A** male academic, 50s

**B** female publisher, c. 60

## Texts S.9: Telephone conversations between disparates

Text S.9.1 (1975-6): Telephone conversations LLC:o

S.9.1a

**B** male academic

**C** female university secretary, 20s

S.9.1b

**A** male university administrator, 40s

**B** same speaker as **C** in S.9.1a

S.9.1c

**D** same speaker as **A** in S.9.1b

**B** same speaker as **C** in S.9.1a and **B** in S.9.1b

S.9.1d

**A** female university researcher, c. 25

**B** female (mother of friend of **A**)

S.9.1e

**B** female university secretary, 20s

**C** male publisher

S.9.1f

**A** male solicitor's clerk

**B** female university secretary

S.9.1g

**A** student

**B** female university lecturer, c. 30

S.9.1h

**A** female university researcher, 40s

**B** male publisher, 30s

S.9.1i

**A** male publisher

**B** female university lecturer, 40s

S.9.1j

**A** female university secretary

**B** female university secretary

S.9.1k

**A** male

**B** female university secretary, 20s

S.9.1l

**A** female producer

**B** female university lecturer, 40s

S.9.1m

**A** female (wife of **C**)

**B** female university secretary, 20s

**C** male university lecturer, 50s?

S.9.1n

**A** female future student

**B** female university secretary

Text S.9.2 (1975): Telephone conversations LLC:o

S.9.2a

**A** male university lecturer, 50s

**X** non-native speaker (not analysed)

S.9.2b

**A** female official

**B** female university secretary, 20s

S.9.2c

**A** female university secretary

**B** future student

S.9.2d

**A** female office secretary

**B** male university researcher, 60s

S.9.2e

**A** male research assistant, late 20s

**b** female university secretary

**C** female university secretary

S.9.2f

**A** female university lecturer, c. 30

**B** male teacher

S.9.2g

**A** same speaker as **B** in S.9.2f

**B** same speaker as **A** in S.9.2f

S.9.2h

**A** female university secretary, 20s

**B** male

S.9.2i

**A** female official

**B** female university secretary, 20s

S.9.2j

**A** female university secretary, 20s

**B** female research student

S.9.2k

**A** female official

**B** female university secretary, 20s

S.9.2l

**A** male academic, 50s

**B** female university secretary, 20s

**C** female publisher's secretary

**D** male publisher

**Texts S.9.3-5: Telephone and dictaphone**

Text S.9.3 (1975): Ansaphone recordings of 63 males and females, mostly academics and secretaries — LLC:o

Text S.9.4 (1985): Radio phone-in — LLC:s

**a** male presenter, c. 50

**b** male financial advisor, c. 25

**c** various participants

Text S.9.5 (1985): Dictaphone LLC:s

- **a** female publisher, 43
- **b** male publisher, 37
- **c** female senior editor, 30
- **d** male publisher, 36

**Texts S.10: Spontaneous commentary, mainly radio**

Text S.10.1 (1964): Cricket LLC:o

- **a** male broadcaster, 50
- **ra** male broadcaster, 63
- **w** male broadcaster, 52
- **y** male broadcaster, 49

Text S.10.2 (1971): Football LLC:o

- **a** male broadcaster
- **b** male broadcaster

Text S.10.3 (1960): Boxing LLC:o

- **a** male sports broadcaster, 38
- **b** male sports commentator, 67
- **c** male referee
- **d** male

Text S.10.4 (1960): Horse racing LLC:o

S.10.4a

- **a** male sports broadcaster, 42

S.10.4b

- **a** same speaker as **a** in S.10.4a
- **b** male racing commentator

S.10.4c

- **a** same speaker as **a** in S.10.4a, b
- **b** same speaker as **b** in S.10.4b

S.10.4d

- **b** same speaker as **b** in S.10.4b

Text S.10.5 (1965): State funeral LLC:o

- **a** male broadcaster, 33
- **b** female broadcaster, c. 45
- **c** male broadcaster, 45
- **d** male broadcaster, 57
- **e** male broadcaster, 57

Text S.10.6 (1973): Royal wedding LLC:o

S.10.6a: TV commentary

- **a** male broadcaster
- **b** male broadcaster, 61

S.10.6b: Radio commentary (on the same occasion as in S.10.6a)

- **a** male broadcaster
- **b** female broadcaster, c. 53 (same speaker as **b** in S.10.5)
- **c** male broadcaster, 53 (same speaker as **c** in S.10.5)

Text S.10.7 LLC:o

S.10.7a (1960): Launching of a submarine

- **a** male broadcaster, 28 (same speaker as **a** in S.10.5)
- **b** male

S.10.7.b (1960): State visit

- **a** female broadcaster, c. 40 (same speaker as **b** in S.10.5)
- **b** male broadcaster, 28 (same speaker as **a** in S.10.5)
- **c** male soldier
- **d** male soldier

S.10.7c (1976): Physics demonstration

- **a** male academic, 44

Text S.10.8 (1976) LLC:o

S.10.8a: Wild life

- **a** male commentator
- **b** male naturalist

S.10.8b: Physics demonstration

- **a** male academic, 44 (same speaker as **a** in S.10.7c)

Text S.10.9 (1976): Science LLC:s

S.10.9a Physics demonstration (from video tape)

- **a** male academic, 44

S.10.9b Biology demonstration

- **a** male academic, 30s

Text S.10.10 (1984): Tennis LLC:s

- **a** male commentator, c. 40
- **b** female commentator, c. 40
- **c** male umpire, c. 50
- **d** male commentator, c. 30
- **e** linesmen & lineswomen, c. 30-60
- **f-l** male commentators (**f** 50, **g** 64, **h** & **j** same speaker, late 20s, **k** c. 40, **l** c. 30 (Australian))
- **m** umpire

Text S.10.11 (1986): Cookery demonstration LLC:s

S.10.11a

- **a** female cookery expert, c. 30
- **b** male vegetarian, c. 30
- **c** male vegetarian, mid 30s

S.10.11b

- **a** female cookery expert, c. 50

**Texts S.11 Spontaneous oration**

Text S.11.1 (1967): Legal cross-examination LLC:o

- **a** male counsel, 65
- **b** male plaintiff, c. 50
- **c** male judge, 64
- **d** male counsel, 26

Text S.11.2 (1974): Dinner speech LLC:o

- **a** male academic, 58

Text S.11.3 (1961-1962): Word game LLC:o

S.11.3a

- **a** male broadcaster and scriptwriter, c. 41

S.11.3b

- **b** male broadcaster and scriptwriter, c. 39

S.11.3c

- **c** same speaker as **b** in S.11.3b

S.11.3d

- **d** same speaker as **b** in S.11.3b and **c** in S.11.3c

S.11.3e

- **e** same speaker as **b** in S.11.3.b, **c** in S.11.3c and **d** in S.11.3d

S.11.3f

**f** same speaker as **a** in S.11.3a

S.11.3g

**g** same speaker as **a** in S.11.3a and **f** in S.11.3f

Text S.11.4 (1975): House of Commons (Question Time) LLC:o

**a1-a3, q1-q28** politicians, all male except **q3**, between 32 and 63 years of age

Text S.11.5 (1975): House of Commons (debate, mostly monologue) LLC:o

**a-h** male politicians, ages 33-66

Text S.11.6 (1986): House of Lords debate LLC:s

**a-g** Lords aged 45-87

(**a** 61, **b** 78, **c** 87, **d** 70, **e** 70, **f** 45, **g** 76)

**Texts S.12: Prepared oration**

Text S.12.1 (1965): Sermons LLC:o

S.12.1a

**a** male minister of religion, c. 36

S.12.1b

**a** male minister of religion, c. 75

S.12.1c

**a** male minister of religion, c. 35

S.12.1d

**a** male minister of religion, c. 45

Text S.12.2: University lectures LLC:o

S.12.2a (1965) Radio lecture

**a** male professor, 54

S.12.2b (1967) Lunch lecture

**a** male academic, 39

Text S.12.3 (1966): Cases in court (slightly interactive) LLC:o

**a** male counsel, 57

**b** male judge, 66

Text S.12.4 (c. 1966): Cases in court (monologues) LLC:o

S.12.4a

**a** male judge, c. 55

S.12.4b

**b** male judge, 64

Text S.12.5 (1972): Political speech (monologue) LLC:o

**a** male politician, 56

Text S.12.6 (1972): Popular lecture (monologue) LLC:o

**a** male builder, c. 70

Text S.12.7 (1983): Foundation oration (monologue) LLC:s

**a** male academic, 49

APPENDIX 2

# Publications using Survey material


Aarts, F. 1971. On the distribution of noun-phrase types in English clause structure. *Lingua* 26, 281-293.

Aarts, J. & W. Meijs (eds). 1984. *Corpus linguistics: Recent developments in the use of computer corpora in English language research.* Amsterdam: Rodopi.

Aarts, J. & W. Meijs (eds). 1986. *Corpus linguistics II: New studies in the analysis and exploitation of computer corpora.* Amsterdam: Rodopi.

Abberton, E. 1977. Nominal group premodification structures. In Bald & Ilson 1977, 29-72.

Aijmer, K. 1984. *Sort of* and *kind of* in English conversation. *Studia Linguistica* 38, 118-128.

Aijmer, K. 1984. *Go to* and *will* in spoken English. In Ringbom & Rissanen 1984, 141-157.

Aijmer, K. 1985. 'Just'. In Bäckman & Kjellmer 1985, 1-11.

Aijmer, K. 1985. What happens at the end of our utterances? The use of utterance-final tags introduced by *and* and *or*. *Papers from the Eighth Scandinavian Conference of Linguistics*, ed O.Togeby, 117-127. Institut for Nordisk Filologi, University of Copenhagen.

Aijmer, K. 1986. Why is *actually* so frequent in spoken English? In Tottie & Bäcklund 1986, 119-129.

Aijmer, K. 1986. Discourse variation and hedging. In Aarts & Meijs 1986, 1-18.

Aijmer, K. 1986. Speaking with many voices: Direct and indirect speech in English conversation. *Papers from the 9th Scandinavian Conference of Linguistics*, ed Ö. Dahl, 1-14. Department of Linguistics, Stockholm University.

Aijmer, K. 1987. *Oh* and *ah* in English conversation. In Meijs 1987, 61-86.

Aijmer, K. 1988. 'Now may we have a word on this': The use of *now* as a discourse particle. In Kytö, Ihalainen & Rissanen 1988, 15-34.

Akimoto, M. 1983. *Idiomaticity*. Tokyo: Shinozaki Shorin.

Akimoto, M. 1989. *A study of verbo-nominal structures in English.* Tokyo: Shinozaki Shorin.

Algeo, J. 1988. British and American grammatical differences. *International Journal of Lexicography* 1, 1-31.

Altenberg, B. 1984. Causal linking in spoken and written English. *Studia Linguistica* 38, 20-69.

Altenberg, B. 1984. Lexical and sex-related differences in spoken and written English: Some results of undergraduate research at Lund University. In Ringbom & Rissanen 1984, 279-298.

Altenberg, B. 1986. Contrastive linking in spoken and written English. In Tottie & Bäcklund 1986, 13-40.

Altenberg, B. 1987. *Prosodic patterns in spoken English. Studies in the correlation between prosody and grammar for text-to-speech conversion.* Lund Studies in English 76. Lund: Lund University Press.

Altenberg, B. 1987. Predicting text segmentation into tone units. In Meijs 1987, 49-60.

Altenberg, B. 1987. Some functions of the booster in spoken English. In Lindblad & Ljung 1987, 5-20.

Altenberg, B. 1987. Causal ordering strategies in English conversation. *Grammar in the construction of texts*, ed J. Monaghan 1987, 50-64. London: Frances Pinter.

Altenberg, B. 1988. Spoken English and the lexicon. In Liedloff 1988, 19-32.

Altenberg, B. & G. Tottie. 1984. Will there be texts in this class? Writing term papers within a research project. In Ringbom & Rissanen 1984, 265-278.

André, E. 1974. Studies in the correspondence between English intonation and the noun phrase in English grammar. Université de Liège.

André, E. 1975. English tone contrasts in relative and adverbial clauses and in enumerative statements. *Revue de Phonétique Appliquée* 35, 107-118.

Bäcklund, I. 1984. *Conjunction-headed abbreviated clauses in English.* Uppsala: Almqvist & Wiksell.

Bäcklund, I. 1986. *Beat until stiff.* Conjunction-headed abbreviated clauses in spoken and written English. In Tottie & Bäcklund 1986, 41-55.

Bäcklund, I. 1989. Cues to the audience. On some structural markers in English monologue. In Odenstedt & Persson 1989, 29-39.

Bäckman, S. & G. Kjellmer (eds). 1985. *Papers on language and literature presented to Alvar Ellegård and Erik Frykman*, Gothenburg Studies in English 60. Gothenburg: Acta Universitatis Gothoburgensis.

Bald, W-D. 1971. The scope of negation and copula sentences in English. *Journal of English Linguistics* 5, 1-28.

Bald, W-D. 1972. *Studien zu den kopulativen Verben des Englischen.* Munich: Hueber.

Bald, W-D. 1979. English tag-questions and intonation. *Anglistentag 1979*, ed K. Schumann, 263-291. Berlin: Technische Universität.

Bald, W-D. 1980. Some functions of *yes* and *no* in conversation. In Greenbaum, Leech & Svartvik 1980, 178-191.

Bald, W-D. 1980. Propword *one*. *Der fremdsprachliche Unterricht* 54, 135-139.

Bald, W-D. 1984. Markedness and diachrony. *Tagungsberichte des Anglistentags*, eds B. Carstensen & H. Grabes, 240-247. Giessen: Hoffman.

Bald, W-D. 1987. *If*-Sätze im Englischen. *Kernprobleme der englischen Grammatik*, ed W-D. Bald, 38-50. Berlin: Langenscheidt-Longman.

Bald, W-D. 1988. A note on the textual distribution of *one*. In Klegraf & Nehls 1988, 156-160.

Bald, W-D. & R. Ilson (eds). 1977. *Studies in English usage: The resources of a present-day English corpus for linguistic analysis*. Frankfurt: Lang.

Bergenholz, H. & B. Schaeder (eds). 1979. *Empirische Textwissenschaft: Aufbau und Auswertung von Text-Corpora*. Königstein: Scriptor.

Biber, D. 1985. Investigating macroscopic textual variation through multi-feature/multi-dimensional analyses. *Linguistics* 23, 155-178.

Biber, D. 1986. Spoken and written textual dimensions in English: Resolving the contradictory findings. *Language* 62, 384-414.

Biber, D. 1986. On the investigation of spoken/written differences. *Studia Linguistica* 40, 1-21.

Biber, D. 1987. A textual comparison of British and American writing. *American Speech* 62, 99-119.

Biber, D. 1988. *Variation across speech and writing*. Cambridge: Cambridge University Press.

Biber, D. & E. Finegan. 1986. An initial typology of English text types. In Aarts & Meijs 1986, 19-46.

Biber, D. & E. Finegan. 1988. Adverbial stance types in English. *Discourse Processes* 11, 1-34.

Biber, D. & E. Finegan. 1988. Drift in three English genres from the 18th to the 20th centuries: A multidimensional approach. In Kytö, Ihalainen & Rissanen 1988, 83-101.

Biber, D. & E. Finegan. 1989. Styles of stance in English. Lexical and grammatical marking of evidentiality and affect. *Text* 9, 93-124.

Black, M. 1977. An investigation into factors influencing the choice between the syllabic and contracted form of *is*. In Bald & Ilson 1977, 171-182.

Boardman, G.M. 1977. A study of certain kinds of anacolutha in a corpus of spoken English. In Bald & Ilson 1977, 183-221.

Breivik, L.E. 1981. On the interpretation of existential *there*. *Language* 57, 1-25.

Breivik, L.E. 1983. *Existential 'there', a synchronic and diachronic study*. Bergen: Department of English, University of Bergen.

Breivik, L.E. 1986. Some remarks on cleft sentences in present-day English. *Linguistics across historical and geographical boundaries*, eds D. Kastovsky & A. Szwedek, vol 2, 815-826. Berlin: Mouton de Gruyter.

Brodda, B. 1988. Tracing turns in the London-Lund Corpus with BetaText. *Literary and Linguistic Computing* 3, 94-104.

Bublitz, W. 1978. *Ausdrucksweisen der Sprechereinstellung im Deutschen und Englischen*. Tübingen: Niemeyer.

Bublitz, W. 1979. Tag questions, transformational grammar, and pragmatics. *Papers and Studies in Contrastive Linguistics* 9, 5-21.

Bublitz, W. 1986. Gesprächsthema und thematische Handlungen im Englischen. *Pragmantax*, eds A. Burkhardt & K-H. Körner, 225-234. Tübingen: Niemeyer.

Bublitz, W. 1988. *Supportive fellow-speakers and cooperative conversations.* Amsterdam: John Benjamins.

Bublitz, W. & H. Weber. 1986. 'I mustn't go on boringly talking about me': Die Behandlung von Alltagsgesprächen im Englischunterricht. *Neusprachliche Mitteilungen aus Wissenschaft und Praxis* 4, 222-231.

Burton-Roberts, N. 1975. Nominal apposition. *Foundations of Language* 13, 391-419.

Burton-Roberts, N. 1976. On the generic indefinite article. *Language* 52, 427-448.

Carvell, H.T. & J. Svartvik. 1969. *Computational experiments in grammatical classification.* The Hague: Mouton.

Chalker, S. 1985. 'Like I said'. *World Englishes* 4, 257-260.

Cheepen, C. 1988. *The predictability of informal conversation.* London: Frances Pinter.

Close, R.A. 1977. Some observations on the meaning and function of verb phrases having future reference. In Bald & Ilson 1977, 125-156.

Close, R.A. 1980. *Will* in *if*-clauses. In Greenbaum, Leech & Svartvik 1980, 100-109.

Close, R.A. 1987. Notes on the split infinitive. *Journal of English Linguistics* 20, 217-229.

Coates, J. 1977. A corpus study of modifiers in sequence. In Bald & Ilson 1977, 9-27.

Coates, J. 1980. On the non-equivalence of *may* and *can*. *Lingua* 50, 209-220.

Coates, J. 1982. *The semantics of the modal auxiliaries.* London: Croom Helm.

Coates, J. 1987. Epistemic modality and spoken discourse. *Transactions of the Philological Society*, 110-131.

Collins, P. 1982. Cleft sentences in English discourse. *Australian Review of Applied Linguistics* 5, 60-83.

Collins, P. 1985. *Th*-clefts and *all*-clefts. *Beiträge zur Phonetik und Linguistik* 48, 45-53.

Collins, P. 1987. Cleft & pseudo-cleft constructions in English spoken and written discourse. *ICAME Journal* 11, 5-17.

Crystal, D. 1969. *Prosodic systems and intonation in English.* Cambridge: Cambridge University Press.

Crystal, D. 1972. The intonation system of English. *Intonation*, ed D. Bolinger, 110-136. Harmondsworth: Penguin.

Crystal, D. 1975. *The English tone of voice. Essays on intonation, prosody and paralanguage.* London: Edward Arnold.

Crystal, D. 1980. Neglected grammatical factors in conversational English. In Greenbaum, Leech & Svartvik 1980, 153-166.

Crystal, D. & D. Davy. 1969. *Investigating English style.* London: Longman.

Crystal, D. & D. Davy. 1975. *Advanced conversational English.* London: Longman.

Crystal, D. & R. Quirk. 1964. *Systems of prosodic and paralinguistic features in English.* The Hague: Mouton.

Davison, B.D. 1977. Aspects of information structuring in modern spoken English. In Bald & Ilson 1977, 157-170.

Eeg-Olofsson, M. 1985. A probability model for computer-aided word-class determination. *ALLC Journal* 5, 25-30.

Eeg-Olofsson, M. 1987. Assigning new tags to old texts - an experiment in automatic word class tagging. In Meijs 1987, 45-47.

Eeg-Olofsson, M. & J. Svartvik. 1984. Four-level tagging of spoken English. In Aarts & Meijs 1984, 53-64.

Engels, L. 1988. The effect of spoken and written-to-be-spoken English on word frequency counts of written English. In Klegraf & Nehls 1988, 407-425.

Enkvist, N.E. (ed). 1982. *Impromptu speech: A symposium.* Åbo: Åbo Akademi.

Erman, B. 1986. Some pragmatic expressions in English conversation. In Tottie & Bäcklund 1986, 131-147.

Erman, B. 1987. *Pragmatic expressions in English: A study of 'you know', 'you see' and 'I mean' in face-to-face conversation.* Stockholm: Almqvist & Wiksell.

Fill, A.F. 1986. 'Divided illocution' in conversational and other situations - and some of its implications. *IRAL* 24, 27-34.

Finegan, E. & D. Biber. 1986. Toward a unified model of sociolinguistic prestige. *Diversity and diachrony*, ed. D. Sankoff, 391-397. Amsterdam: John Benjamins.

Fjelkestam-Nilsson, B. 1983. *'Also' and 'too': A corpus-based study of their frequency and use in modern English.* Stockholm: Almqvist & Wiksell.

Forsheden, O. 1983. Studies on contraction in the London-Lund Corpus of Spoken English. Department of English, Lund University.

Garnham, A., R.C. Shillcock, G.D.A. Brown, A.I.D. Mill & A. Cutler. 1981. Slips of the tongue in the London-Lund corpus of spontaneous conversation. *Linguistics* 19, 805-817. [Reprinted in *Slips of the tongue and language production*, ed A. Cutler, 251-263. The Hague: Mouton, 1982.]

Geluykens, R. 1984. *Focus phenomena in English: An empirical investigation into cleft and pseudo-cleft sentences.* Antwerp Papers in Linguistics 36. University of Antwerp.

Geluykens, R. 1986. *Questioning intonation: An empirical study into the prosodic feature 'rising intonation' and its relevance for the production and recognition of questions.* Antwerp Papers in Linguistics 48. University of Antwerp.

Geluykens, R. 1987. Tails (right-dislocations) as a repair mechanism in English conversation. *Getting one's words into line: On word order and functional grammar*, eds J. Nuyts and G. de Schutter, 119-129. Dordrecht: Fortis Publications.

Geluykens, R. 1987. Intonation and speech act type: An experimental approach to rising intonation in queclaratives. *Journal of Pragmatics* 11, 483-494.

Geluykens, R. 1988. On the myth of rising intonation in polar questions. *Journal of Pragmatics* 12, 467-485.

Gnutzmann, C. 1984. Textkonstitution und Textsortendifferenzierung. Zum Gebrauch der englischen Demonstrativpronomina. *Arbeiten aus Anglistik und Amerikanistik* 9, 107-130.

Gnutzmann, C., R. Ilson & J. Webster. 1973. Comparative constructions in contemporary English. *English Studies* 54, 417-438.

Granger, S. 1983. *The 'be' + past participle construction in spoken English with special emphasis on the passive.* Amsterdam: North-Holland.

Greenbaum, S. 1969. *Studies in English adverbial usage.* London: Longman.

Greenbaum, S. 1984. Corpus analysis and elicitation tests. In Aarts & Meijs 1984, 193-201.

Greenbaum, S. 1985. 25th anniversary of the Survey of English Usage. *World Englishes* 4, 261-265.

Greenbaum, S. 1988. Syntactic devices for compression in English. In Klegraf & Nehls 1988, 3-10.

Greenbaum, S. 1988. *Good English and the grammarian.* London: Longman.

Greenbaum, S., G. Leech & J. Svartvik (eds). 1980. *Studies in English linguistics for Randolph Quirk.* London: Longman.

Haan, P. de. 1984. Relative clauses compared. *ICAME News* 8, 47-59.

Haase, I. 1988. Temporafolge und Bedingungstyp im konditionalen Satzgefüge. In Mindt 1988, 69-76.

Haegeman, L. 1980. *Have to* and progressive aspect. *Journal of English Linguistics* 14, 1-5.

Haegeman, L. 1981. Modal *shall* and speaker's control. *Journal of English Linguistics* 15, 4-9.

Haegeman, L. 1983. *The semantics of 'will' in present-day English: A unified account.* Brussels: AWLK.

Haegeman, L. 1984. Pragmatic conditionals in English. *Folia Linguistica* 13, 485-502.

Hedström, K. 1984. A study of repairs in speech. Stockholm Papers in English Language and Literature 4, 69-101. Department of English, Stockholm University.

Hermerén, L. 1986. Modalities in spoken and written English: An inventory of forms. In Tottie & Bäcklund 1986, 57-91.

Hiller, U. 1984. *Untersuchungen zu kontextuellen und registerspezifischen Einflüssen auf Morphemreduktionen im Englischen.* Frankfurt: Peter Lang.

Hiller, U. 1987. *She isn't studying* v. *She's not studying*: An investigation into the choice between the two contracted variants of negated English auxiliaries. *Die Neueren Sprachen* 86, 531-553.

Holmes, J. 1988. Doubt and certainty in ESL textbooks. *Applied Linguistics* 9, 21-44.

Hüllen, W. 1986. New aspects for foreign language learning and teaching from conversational analysis. *Linguistics across historical and geographical boundaries*, eds D. Kastovsky & A. Szwedek, vol 2: 1219-1232.

*ICAME News*. 1978-1986. Newsletter of the International Computer Archive of Modern English. Bergen: Norwegian Computing Centre for the Humanities.

*ICAME Journal.* 1987- . Journal of the International Computer Archive of Modern English. Bergen: Norwegian Computing Centre for the Humanities.

Ikegami, Y. 1988. '*Have* + object + past participle' and '*get* + object + past participle' in the SEU Corpus. *Meaning and beyond*, eds U. Fries & M. Heusser, 197-223. Tübingen: Gunter Narr.

Johansson, S. 1982. Review of Svartvik & Quirk 1980. *Computers and the Humanities* 16, 174-180.

Johansson, S. 1987. Machine-readable texts in English language research: Progress and prospects. In Lindblad & Ljung 1987, 125-137.

Johannesson, N-L. 1982. On the use of post-modification in English noun phrases. *The Eighth LACUS Forum 1981*, eds W. Gutwinski & G. Jolly, 187-195. Columbia, S.C.: Hornbeam Press.

Kaye, G. 1988. The design of the database for the Survey of English Usage. In Kytö, Ihalainen & Rissanen 1988, 145-168.

Klegraf, J. & D. Nehls (eds). 1988. *Essays on the English language and applied linguistics on the occasion of Gerhard Nickel's 60th Birthday.* Heidelberg: Julius Groos.

Kohda, M. 1985. 'The stone is too heavy for me to lift (*it)'. *Eibei Kenkyu* 14, 47-72.

Krenn, M. 1985. *Probleme der Diskursanalyse im Englischen. Verweise mit 'this', 'that', 'it' und Verwandtes.* Tübingen: Gunter Narr.

Krenn, M. 1985. Proformen und Diskursorganisation. *Grammatik, Semantik, Textlinguistik*, eds W. Kürschner & R. Vogt, 355-362. Tübingen: Niemeyer.

Kussmaul, P. 1977. Auffordern im Englischen. *Kongressberichte der 8. Jahrestagung der GAL*, 57-66.

Kussmaul, P. 1978. *In fact, actually, anyway ...* : Indikatoren von Sprechakten im informellen gesprochenen Englisch. *Die Neueren Sprachen* 27, 357-369.

Kytö, M., O. Ihalainen & M. Rissanen (eds). 1988. *Corpus linguistics, hard and soft.* Amsterdam: Rodopi.

Leech, G., R. Garside & E.S. Atwell. 1983. Recent developments in the use of computer corpora in English language research. *Transactions of the Philological Society*, 32-40.

Leech, G. & J. Svartvik. 1975. *A communicative grammar of English.* London: Longman.

Liedloff, V. (ed). 1988. *Standardization in computerized lexicography. Proceedings of the IV Forum for Information Science and Practice.* Saarbrücken: IAI.

Lindblad, I. & M. Ljung (eds). 1987. *Proceedings from the Third Nordic Conference for English Studies*, vol I. Stockholm Studies in English 73. Stockholm: Almqvist & Wiksell.

Lipka, L. 1972. *Semantic structure and word-formation. Verb-particle constructions in contemporary English.* Munich: Fink.

Mair, C. 1987. Tough-movement in present-day British English: A corpus-based study. *Studia Linguistica* 41, 59-71.

Mair, C. 1987. Instabile Infinitivkonstruktionen im heutigen Englisch. *Linguistische Berichte* 3, 381-397.

Mair, C. 1987. *For/to*-infinitival clauses in contemporary British English: A study based on the material collected in the Survey of English Usage, University College London. *English Studies* 68, 545-559.

Mair, C. 1988. Extraposed gerundial subject clauses in present-day British English. An investigation of the corpus of the Survey of English Usage (University College London). *Arbeiten aus Anglistik und Amerikanistik* 13, 51-63.

Mair, C. 1990. *Infinitival complement clauses in English: A study of syntax in discourse.* Cambridge: Cambridge University Press.

McCarthy, M. 1988. Some vocabulary patterns in conversation. *Vocabulary and language teaching*, eds R. Carter & M. McCarthy, 181-200. London: Longman.

Meijs, W. (ed). 1987. *Corpus linguistics and beyond.* Amsterdam: Rodopi.

Meyer, C. 1987. Apposition in English. *Journal of English Linguistics* 20, 101-121.

Michiels, A. 1977. Relative pronouns in time adjuncts. *English Studies* 58, 26-37.

Mindt, D. 1986. Futurisches *shall* für den Hauptschüler? *1945-85: Vierzig Jahre Englischunterricht für alle*, eds I. Harks-Hanke & W. Zydatiss, 140-151. Berlin: Cornelsen-Velhagen & Klasing.

Mindt, D. (ed). 1988. *EDV in der Angewandten Linguistik: Ziele - Methoden - Ergebnisse.* Frankfurt/Main: Diesterweg.

Mindt, D. 1988. Rechnergestützte Forschungen zur Angewandten Linguistik. In Mindt (ed) 1988, 7-19.

Mindt, D. & I. Haase. 1988. Definition linguistischer Daten und Modifikation durch Variablenänderung: das englische Verbsystem. In Mindt (ed) 1988, 45-53.

Nässlin, S. 1984. *The English tag question: A study of sentences containing tags of the type 'isn't it?', 'is it?'*. Stockholm: Almqvist & Wiksell.

Nehls, D. 1986. *Semantik und Syntax des englischen Verbs, Teil II. Die Modalverben: Eine kontrastive Analyse der Modalverben im Englischen und Deutschen.* Heidelberg: Julius Groos.

Nevalainen, T. 1987. Adverbial focusing and intonation. *Lingua* 73, 141-165.

Nevalainen, T. 1987. The rhythm hypothesis of adverb placement: A case study of *only. Neuphilologische Mitteilungen* 88, 365-377.

Nevalainen, T. 1987. A quantitative survey of nuclear tone variation in English. *Proceedings of the Eleventh International Congress of Phonetic Sciences*, vol 3: 180-183. Tallin: Institute of Language and Literature, Academy of Sciences of the Estonian S.S.R.

Nevalainen, T. & M. Rissanen. 1986. 'Do you support the *do*-support?' Emphatic and non-emphatic *do* in affirmative statements in present-day spoken English. *Papers from the Third Scandinavian Symposium on Syntactic Variation*, ed S. Jacobson, 35-50. Stockholm: Almqvist & Wiksell.

Odenstedt, B. & G. Persson (eds). 1989. *Instead of flowers. Papers in honour of Mats Rydén on the occasion of his sixtieth birthday, August 27, 1989.* Acta Universitatis Umensis 90. Stockholm: Almqvist & Wiksell.

Oostdijk, N. 1988. A corpus linguistic approach to linguistic variation. *Literary and Linguistic Computing* 3, 12-25.

Oreström, B. 1982. When is it my turn to speak? In Enkvist 1982, 267-276.

Oreström, B. 1983. *Turn-taking in English conversation.* Lund Studies in English 66. Lund: Lund University Press.

Pakosz, M. 1988. Tonic prominence and the coding of thematic-rhematic relations. *Papers in Contrastive Linguistics* 23, 119-134.

Palmer, F.R. 1979. *Modality and the English modals.* London: Longman.

Palmer, F.R. 1980. *Can, will* and *actually.* In Greenbaum, Leech & Svartvik 1980, 91-99.

Pannek, G. 1988. Relativsätze im gesprochenen Englisch und in Lehrwerken für den Englischunterricht. In Mindt 1988 (ed), 77-83.

Persson, G. 1974. *Repetition in English: Part 1, Sequential repetition.* Uppsala: Uppsala University.

Quirk, R. 1957. Relative clauses in educated spoken English. *English Studies* 38, 97-109. [Reprinted in Quirk 1968, 95-113.]

Quirk, R. 1968. *Essays on the English language: Medieval and modern.* London: Longman.

Quirk, R. 1972. *The English language and images of matter.* Oxford: Oxford University Press.

Quirk, R. 1974. *The linguist and the English language.* London: Edward Arnold.

Quirk, R. 1984. Recent work on adverbial realisation and position. In Aarts & Meijs 1984, 185-192.

Quirk, R. & D. Crystal. 1986. On scales of contrast in connected English speech. *In Memory of J.R. Firth*, eds C.E. Bazell, H.C. Catford, M.A.K. Halliday & R.H. Robins, 359-369. London: Longman. [Reprinted in Quirk 1968, 136-147.]

Quirk, R., A. Duckworth, J. Rusiecki, J. Svartvik & A. Colin. 1964. Studies in the correspondence of prosodic to grammatical features in English. *Proceedings of the Ninth International Congress of Linguists*, 679-691. The Hague: Mouton. [Reprinted in Quirk 1968, 120-135.]

Quirk, R. & S. Greenbaum. 1973. *A university grammar of English.* London: Longman. [= *A concise grammar of contemporary English.* New York: Harcourt, Brace, Jovanovich.]

Quirk, R., S. Greenbaum, G. Leech & J. Svartvik. 1972. *A grammar of contemporary English.* London: Longman.

Quirk, R., S. Greenbaum, G. Leech & J. Svartvik. 1985. *A comprehensive grammar of the English language*. London: Longman.

Quirk, R. & J. Mulholland. 1964. Complex prepositions and related sequences. *English Studies presented to W.W. Zandvoort*. Supplement to *English Studies* 45, 64-73. [Reprinted in Quirk 1968, 148-160.]

Quirk, R. & J. Svartvik. 1979. A corpus of modern English. In Bergenholz & Schaeder 1979, 204-218.

Ringbom, H. & M. Rissanen (eds). 1984. *Proceedings from the Second Nordic Conference for English Studies*. Åbo: Åbo Akademi.

Rissanen, M. 1979. On the position of *only* in present-day written English. *Papers from the Scandinavian Symposium on Syntactic Variation*, ed S. Jacobson, 63-76. Stockholm: Almqvist & Wiksell.

Rusiecki, J. 1985. *Adjectives and comparison in English. A semantic study*. London: Longman.

Rycker, T. de. 1984. Imperative structures: Form and function in conversational English. *Antwerp Papers in Linguistics* 38. Linguistics Department, University of Antwerp.

Sahlin, E. 1979. *'Some' and 'any' in spoken and written English*. Studia Anglistica Upsaliensia 38. Stockholm: Almqvist & Wiksell.

Stenström, A-B. 1982. Feedback. In Enkvist 1982, 319-340.

Stenström, A-B. 1983. Questioning strategies in English and Swedish conversation. *Cross-language analysis and second language acquistion* 2, 67-78. Department of English, University of Jyväskylä.

Stenström, A-B. 1984. *Questions and responses in English conversation*. Lund Studies in English 68. Lund: Lund University Press.

Stenström, A-B. 1984. Discourse tags. In Aarts & Meijs 1984, 65-81.

Stenström, A-B. 1985. English in speech and writing. *Papers and Studies in Contrastive Linguistics*, 115-130. Poznan: Adam Mickiewicz University.

Stenström, A-B. 1986. What does *really* really do? Strategies in speech and writing. In Tottie & Bäcklund 1986, 149-163. [Also in *Grammar in the construction of texts*, ed J. Monaghan, 65-79. London: Frances Pinter, 1987.]

Stenström, A-B. 1986. A study of pauses as demarcators in discourse and syntax. In Aarts & Meijs 1986, 203-218.

Stenström, A-B. 1987. Carry-on signals in English conversation. In Meijs 1987, 87-119.

Stenström, A-B. 1988. Adverbial commas and prosodic segmentation. In Kytö, Ihalainen & Rissanen 1988, 221-236.

Stenström, A-B. 1988. Questioning in conversation. *Questions and questioning*, ed M. Meyer, 304-325. Berlin: Walter de Gruyter.

Svartvik, J. 1966. *On voice in the English verb*. The Hague: Mouton.

Svartvik, J. 1980. *Well* in conversation. In Greenbaum, Leech & Svartvik 1980, 167-177.

Svartvik, J. 1980. Tagging spoken English. *ALVAR: A linguistically varied assortment of readings. Studies presented to Alvar Ellegård on the occasion of his 60th birthday*, eds J. Allwood & M. Ljung, 182-206. Department of English, Stockholm University.

Svartvik, J. 1982. The segmentation of impromptu speech. In Enkvist 1982, 131-145.

Svartvik, J. 1987. Graphic English prosody. In Lindblad & Ljung 1987, 335-346.

Svartvik, J. 1987. Taking a new look at word class tags. In Meijs 1987, 33-43.

Svartvik, J. 1988. On the study of spoken English. *Sprache und Individuum. Kongressbeiträge der 17. Jahrestagung der Gesellschaft für Angewandte Linguistik*, eds W. Kühlwein & B. Spillner, 30-37. Tübingen: Gunter Narr.

Svartvik, J. 1988. A computerized corpus of spoken English. In Liedloff 1988, 33-52.

Svartvik, J. & M. Eeg-Olofsson. 1982. Tagging the London-Lund Corpus of Spoken English. *Computer corpora in English language research*, ed S. Johansson, 85-109. Bergen: Norwegian Computing Centre for the Humanities.

Svartvik, J., M. Eeg-Olofsson, O. Forsheden, B. Oreström & C. Thavenius. 1982. *Survey of Spoken English: Report on research 1975-81*. Lund: Lund University Press.

Svartvik, J. & R. Quirk (eds). 1980. *A corpus of English conversation*. Lund Studies in English 56. Lund: Lund University Press.

Svartvik, J. & A-B. Stenström. 1985. Words, words, words: The rest is silence? In Bäckman & Kjellmer 1985, 342-353.

Taglicht, J. 1977. Relative clauses as postmodifiers: Meaning, syntax and intonation. In Bald & Ilson 1977, 73-107.

Taglicht, J. 1983. *Message and emphasis: On focus and scope in English*. London: Longman.

Taglicht, J. 1987. Relative clauses. *Kernprobleme der englischen Grammatik*, ed W-D. Bald, 244-253. Berlin: Langenscheidt-Longman.

Tesch, F. 1988. *Some* und *any* in affirmativen und negativen Kontexten. In Mindt (ed) 1988, 59-68.

Tesch, F. 1989. Einsatz der elektronischen Datenverarbeitung in der Angewandten Linguistik: Auswertungsergebnisse am Beispiel von *some* und *any*. *Sprechen und Hören: Akten des 23. Linguistischen Kolloquiums, Berlin 1988*, ed N. Reiter, 351-358. Tübingen: Niemeyer.

Tesch, F. 1989. *Die Indefinitpronomina 'some' und 'any' im authentlischen englischen Sprachgebrauch und in Lehrwerken: Eine empirische Untersuchung*. Freie Universität Berlin [microfiche].

Thavenius, C. 1982. Exophora in English conversation: A study of third person pronominal reference. In Enkvist 1982, 291-305.

Thavenius, C. 1983. *Referential pronouns in English conversation*. Lund Studies in English 64. Lund: Lund University Press.

Thavenius, C. 1984. Pronominal chains in English conversation. In Ringbom & Rissanen 1984, 209-219.

Tottie, G. 1980. Affixal and non-affixal negation in English: Two systems in (almost) complementary distribution. *Studia Linguistica* 34, 101-123.

Tottie, G. 1981. Negation and discourse strategy in spoken and written English. *Variation omnibus*, ed H. Cedergren & D. Sankoff, 271-284. Edmonton, Alberta: Linguistic Research.

Tottie, G. 1982. Where do negative sentences come from? *Studia Linguistica* 36, 88-105.

Tottie, G. 1983. The missing link?, or, Why is there twice as much negation in spoken English as in written English? *Papers from the Second Scandinavian Symposium on Syntactic Variation*, ed S. Jacobson, 67-74. Stockholm: Almqvist & Wiksell.

Tottie, G. 1984. Is there an adverbial in this text? In Ringbom & Rissanen 1984, 299-315. [Also in *Diversity and diachrony*, ed D. Sankoff, 139-152. Amsterdam: John Benjamins, 1986.]

Tottie, G. 1986. The importance of being adverbial: Adverbials of focusing and contingency in spoken and written English. In Tottie & Bäcklund 1986, 93-118.

Tottie, G. 1988. *No*-negation and *not*-negation in spoken and written English. In Kytö, Ihalainen & Rissanen 1988, 245-265.

Tottie, G. Forthcoming. Lexical diffusion in syntactic change: Frequency as a determinant of linguistic conservatism in the development of negation in English. In *Proceedings from the Conference on Historical English Syntax*, September 12-14, 1988, ed D. Kastovsky. Berlin: Mouton de Gruyter.

Tottie, G., B. Altenberg & L. Hermerén. 1983. English in speech and writing. A manual for students. ETOS Report, Departments of English, Lund and Uppsala Universities.

Tottie, G. & I. Bäcklund (eds). 1986. *English in speech and writing: A symposium.* Stockholm: Almqvist & Wiksell.

Tottie, G., M. Eeg-Olofsson & C. Thavenius. 1984. Tagging negative sentences in LOB and LLC. In Aarts & Meijs 1984, 173-184.

Tottie, G. & C. Paradis. 1982. From function to structure: Some pragmatic determinants of syntactic frequencies in impromptu speech. In Enkvist 1982, 307-317.

Umeda, I. 1987. Psychological predicates in English. *International Review of Applied Linguistics* 25, 91-101.

Vandespitte, S. 1987. Prolegomena to the analysis of *because*-clauses. *Studies in honour of René Derolez*, ed A.M. Simon-Vandenbergen, 636-661. Gent: Seminarie voor Engelse en Oud-Germaanse Taalkunde, R.U.G.

Viethen, H.W. 1979. *Untersuchungen zum Englischen Present Perfect und zum Generischen.* Frankfurt: Peter Lang.

Viitanen, O. 1986. On the position of *only* in English conversation. In Tottie & Bäcklund 1986, 165-175.

Wales, K. 1980. 'Personal' and 'indefinite' reference: The uses of the pronoun *one* in present-day English. *Nottingham Linguistic Circular* 9, 93-117.

Wales, K. 1980. Exophora re-examined: The uses of *we* in present-day English. *UEA Papers in Linguistics* 12, 21-44.

Wales, K. 1984. Problems of pronoun concord, or no one's sexist, are they? *UEA Papers in Linguistics* 20, 28-45.

Wales, K. 1985. Generic *your* and Jacobean drama: The rise and fall of a pronominal usage. *English Studies* 66, 7-24.

Webster, J. 1977. A corpus-based exploration of prepositional phrases postmodifying verbally-related heads. In Bald & Ilson 1977, 109-124.

Yadugiri, M.A. 1986. Some pragmatic implications of the use of *yes* and *no* in response to *yes-no* questions. *Journal of Pragmatics* 10, 199-210.

## Note

We are grateful to Bengt Altenberg for making available his bibliography of uses of computerized corpora, published in *ICAME News* 10 (1986) and Altenberg (forthcoming a).

# Part II: Research

# 2

# The TESS project

*Jan Svartvik*

## 2.1 Aims

The governing idea behind the research project Text Segmentation for Speech (with the acronym TESS) has been to make good use of the London-Lund Corpus of Spoken English after the original version (LLC:o) had become available in computerized form, as described in Chapter 1. Our aim was, first, to contribute to a description of some of the linguistic rules that govern the prosodic segmentation of natural English speech and, second, to 'reverse' these rules into algorithms which could be used for improving automatic text-to-speech conversion. The approach was based on the notion that an in-depth study of grammatical, lexical, pragmatic and prosodic properties in a large corpus of authentic spoken language like LLC would provide information of value for making predictions about prosodic segmentation, and that such insights would prove useful in, for example, work on producing synthetic speech.

## 2.2 Sponsors, staff and publications

The TESS project has been carried out in the Survey of Spoken English, Department of English, Lund University, with the Bank of Sweden Tercentenary Foundation as its main sponsor. The staff consisted of Jan Svartvik (director), Bengt Altenberg (assistant director), Mats Eeg-Olofsson, and Anna-Brita Stenström. From the beginning of the project, members of the team have reported on their research in papers read at various scholarly events, in particular at the ICAME conferences, organized under the auspices of the International Computer Archive of Modern English (Lancaster 1984, Lund 1985, Amsterdam 1986, Helsinki 1987 and Birmingham 1988). A list of bibliographical references to publications relating to the TESS project can be seen in Chapter 1, Appendix 2, under the names of the staff members. Further documentation of the project is provided in subsequent chapters of this book.

As the only member of the TESS team working full time on the project, Bengt Altenberg has carried out the main part of the research (see, in particular, his monograph Altenberg 1987a and numerous other publications listed in Appendix 2). His contributions also make up the major part of this volume: 'Spoken English and the dictionary' (Chapter 6), 'Some functions of the booster' (Chapter 7), 'Predicting text segmentation into tone units' (Chapter 11), and 'Automatic text segmentation into tone units' (Chapter 12).

In addition to numerous contributions in previous publications, as can be seen in the bibliographical references, Anna-Brita Stenström has submitted three chapters to this volume: 'Lexical items peculiar to spoken discourse' (Chapter 5), 'Pauses in monologue and dialogue' (Chapter 8), and 'Adverbial commas and prosodic segmentation' (Chapter 9).

Mats Eeg-Olofsson has written our computer programs and also two chapters: 'An automatic word-class tagger and a phrase parser' (Chapter 4) and 'A Prolog implementation of automatic segmentation' (Chapter 13).

My own contributions to this volume, apart from editing it, consist of the introductory Chapter 1, 'The London-Lund Corpus of Spoken English', co-authored with Sidney Greenbaum, the present chapter (2), 'Tagging and parsing on the TESS project' (Chapter 3), and 'Graphic English prosody' (Chapter 10).

## 2.3 Tie-in with previous research

The TESS project has been carried out within the general framework of the Survey of Spoken English, and it will therefore be relevant to begin this account by linking up with previous local research.

While the primary aim of the Survey, when it was launched at Lund in 1975, was to computerize the then available spoken texts of the corpus (as described in Chapter 1), the secondary aim was to make use of this computerized material for research. On the completion of the first objective at the end of the seventies with the production of the computerized tape (LLC:o, see p 14) and the book (Svartvik & Quirk 1980), we were able to shift our attention to the secondary aim and consider ways of making good use of this data collection. The long-term plan of the Survey is to improve our knowledge of authentic spoken English and to provide better descriptions of the types of discourse represented in this corpus.

One of the early results from the second phase was a concordance (available from ICAME, see p 17), which was also used for producing frequency lists. Table 2:1 shows the 64 most frequent lexical items in LLC and, to highlight some contrasting word frequencies in written English, also the rank order of the same items in LOB and Brown. (There are two separate rank listings for LLC: one for text categories S.1-3, ie surreptitiously recorded face-to-face conversations, and another for the remaining text categories S.4-12; all frequencies are based on version LLC:o. For more extensive frequency lists, see Svartvik et al 1982: 41ff.)

Further documentation of LLC-based research has appeared in doctoral theses: by Oreström (1983) on turn-taking, Thavenius (1983) on referential pronouns, and Stenström (1984a) on questions and responses. In addition, there have been corpus-related studies by other colleagues at Lund (see entries under the names of Karin Aijmer, Lars Hermerén and Gunnel Tottie in Chapter 1, Appendix 2). Doctoral theses based on LLC have also appeared in other English departments in Sweden: at Uppsala by Sahlin (1979) on *some* and *any*, and at Stockholm by Fjelkestam-Nilsson (1983) on *also* and *too*, Nässlin (1984) on tag questions, and Erman (1987) on pragmatic expressions.

However, the most important use of the spoken corpus will no doubt be made by the international community of scholars who have acquired copies (at first available only on computer tape, now also on floppy disk). Not only does the use of a computerized corpus give researchers in any part of the world access to material that would otherwise be difficult or impossible for them to obtain (for example private, informal conversation), but such a generally available 'standard' corpus also enables them to benefit from sharing, with colleagues all over the world, a common basis for corpus-based studies.

In the early eighties there began a second departmental project, 'English in Speech and Writing' (ETOS), directed by Gunnel Tottie and Bengt Altenberg and sponsored by the Swedish Council for Research in the Humanities and Social Sciences. Subsequently, after Gunnel Tottie's transfer to Uppsala University, it continued as a joint Lund-Uppsala venture. This project was contrastive in character, comparing certain features in LLC with parallel features in LOB (see for example the collection of articles in Tottie & Bäcklund 1986).

**Table 2:1.** Rank list of the most frequent words in LLC compared with LOB and Brown. A dash denotes that the item is not among the 100 most frequent words.

| WORD | LONDON-LUND CORPUS S.1-3 | LONDON-LUND CORPUS S.4-12 | LOB | BROWN |
|---|---|---|---|---|
| I | 1 | 5 | 17 | 20 |
| the | 2 | 1 | 1 | 1 |
| and | 3 | 2 | 3 | 3 |
| to | 4 | 3 | 4 | 4 |
| you | 5 | 9 | 32 | 33 |
| of | 6 | 4 | 2 | 2 |
| a | 7 | 6 | 5 | 5 |
| it | 8 | 10 | 10 | 12 |
| in | 9 | 8 | 6 | 6 |
| that | 10 | 7 | 7 | 7 |
| yes | 11 | 14 | - | - |
| was | 12 | 12 | 9 | 9 |
| is | 13 | 11 | 8 | 8 |
| well | 14 | 22 | 95 | 100 |
| know | 15 | 28 | - | - |
| but | 16 | 16 | 24 | 25 |
| this | 17 | 13 | 22 | 21 |
| he | 18 | 21 | 12 | 10 |
| they | 19 | 26 | 33 | 30 |
| on | 20 | 15 | 16 | 16 |
| it's | 21 | 27 | - | - |
| [ə:m] | 22 | 33 | - | - |
| have | 23 | 20 | 26 | 28 |
| so | 24 | 32 | 46 | 52 |
| no | 25 | 37 | 47 | 49 |
| think | 25 | 30 | - | - |
| for | 27 | 19 | 11 | 11 |
| be | 28 | 17 | 15 | 17 |
| we | 29 | 18 | 40 | 41 |
| all | 29 | 33 | 39 | 36 |
| [m̀] | 31 | 58 | - | - |
| what | 32 | 41 | 55 | 54 |
| at | 33 | 24 | 19 | 18 |

| WORD | LONDON-LUND CORPUS | | LOB | BROWN |
|---|---|---|---|---|
| | S.1-3 | S.4-12 | | |
| do | 34 | 47 | 73 | 72 |
| very | 35 | 31 | 81 | - |
| one | 36 | 36 | 38 | 32 |
| about | 37 | 42 | 54 | 57 |
| not | 38 | 29 | 23 | 23 |
| don't | 39 | 61 | - | - |
| as | 40 | 23 | 13 | 14 |
| there | 41 | 35 | 36 | 38 |
| with | 42 | 25 | 14 | 13 |
| oh | 43 | 52 | - | - |
| mean | 44 | 89 | - | - |
| got | 45 | 50 | - | - |
| see | 45 | 60 | - | - |
| or | 47 | 43 | 31 | 27 |
| if | 48 | 38 | 45 | 50 |
| are | 49 | 40 | 27 | 24 |
| yeah | 50 | 46 | - | - |
| would | 51 | 56 | 43 | 39 |
| sort | 52 | - | - | - |
| that's | 53 | 72 | - | - |
| just | 54 | 49 | - | - |
| which | 55 | 39 | 28 | 31 |
| I'm | 56 | 65 | - | - |
| had | 57 | 50 | 21 | 22 |
| because | 58 | 63 | - | - |
| really | 59 | 90 | - | - |
| said | 60 | 87 | 52 | 53 |
| then | 61 | 67 | 68 | 71 |
| like | 62 | 76 | 83 | 78 |
| up | 63 | 62 | 56 | 55 |
| when | 64 | 71 | 44 | 45 |
| get | 64 | 76 | - | - |
| my | 66 | 55 | 58 | 75 |
| me | 67 | 68 | 66 | 82 |
| now | 68 | 44 | 72 | 76 |
| from | 69 | 45 | 25 | 26 |

| WORD | LONDON-LUND CORPUS S.1-3 | S.4-12 | LOB | BROWN |
|---|---|---|---|---|
| were | 69 | 58 | 35 | 34 |
| people | 71 | 82 | 100 | - |
| she | 72 | 53 | 30 | 37 |
| them | 73 | 75 | 60 | 59 |
| [əm] | 74 | 73 | - | - |
| out | 75 | 69 | 53 | 51 |
| can | 76 | 66 | 61 | 61 |
| an | 76 | 84 | 34 | 29 |
| thing | 78 | 99 | - | - |
| been | 78 | 64 | 37 | 43 |
| going | 80 | 70 | - | - |
| quite | 81 | - | - | - |
| I've | 82 | - | - | - |
| go | 83 | 94 | - | - |
| he's | 84 | - | - | - |
| [$\vec{m}$] | 84 | - | - | - |
| some | 86 | 95 | 57 | 65 |
| time | 87 | 86 | 63 | 66 |
| say | 88 | 92 | - | - |
| much | 89 | 93 | 89 | 97 |
| did | 90 | - | - | 88 |
| him | 91 | 88 | 49 | 42 |
| two | 91 | 74 | 67 | 69 |
| who | 93 | 79 | 50 | 46 |
| his | 94 | 57 | 18 | 15 |
| good | 95 | - | - | - |
| something | 96 | - | - | - |
| right | 97 | 54 | - | - |
| more | 97 | 97 | 51 | 48 |
| by | 99 | 48 | 20 | 19 |
| any | 99 | - | 75 | 74 |

The major project now in progress is called 'Phraseology in Spoken English', directed by Bengt Altenberg and sponsored by the Bank of Sweden Tercentenary Foundation (see Altenberg forthcoming b, c, d).

## 2.4 Renaissance people needed for speech processing?

The idea of launching the TESS project was very largely triggered off by the impression that the quality of speech synthesis was disappointingly low, an impression which has later been reinforced in statements made by outstanding representatives of speech technology research, for example Jonathan Allen at MIT (1985:1539):

> Speech processing, including speech synthesis, speech recognition, and speaker verification or recognition, continues to provide an exciting example of the design of large, complex computing systems. Despite years of effort and substantial progress, the goal of high-performance systems for these purposes still eludes us, and although commercial systems are available for these purposes, performance approximating that of a human is still a long way off.

It seemed to me that much of the work in the field of artificial speech did not take enough cognizance of the lessons that could be drawn from a study of genuine speech. At the time when our computerized corpus had become available it was therefore a natural step to try and see if this database could provide some good lessons about natural speech to be used for making predictions about prosodic segmentation in text-to-speech conversion.

For the project team it was of course a tall order to enter this new field, and we did not fail to spot several obstacles to making it a successful project: in view of Bolinger's (1972b) claim that accent is predictable only to a mind-reader, the whole project might have been written off as totally misconceived; also, the competence of the members of the research team was chiefly in the areas of grammar, discourse, intonation and computational linguistics, but not in phonetics and speech technology; we did not possess all the necessary equipment for work in speech processing, nor was the recording quality of the corpus designed for such use. The reason why, in the face of such formidable obstacles, we nevertheless decided to launch the project was the following.

As for predictability, while a speaker is of course a free agent in his choice of linguistic behaviour, there is still enough evidence of a connection between grammar and prosody to make it worthwhile to explore more fully in which areas such patterning exists. Lack of competence in the areas required for producing synthetic speech did not worry us too much in view of the fact that the phonetic aspect could confidently be left to professional phoneticians and engineers engaged in such work in both universities and industry. (For two illuminating surveys of speech synthesis, speech perception and related issues, see Allen, Hunnicut & Klatt 1987 and Lieberman & Blumstein 1988.) The problem involving both understanding the human speech process and producing better artificial speech represents such a tremendous challenge that, to hope for any success at all, the attack must be multi-pronged. Clearly, the focus of the TESS contribution was to be on the field where we had our strongest side: access to a large corpus of natural speech and experience of working with it, chiefly from the grammatical end. It is probably significant that Jonathan Allen emphasizes the need for large quantities of material as well as linguistic methodology in future research - and also 'Renaissance people' (1985b:1550):

> Large amounts of speech must be obtained and classified ... It would seem that workers in this field need to become Renaissance people. Indeed, this is so, although the needed breadth of view can be used to develop effective collaborations between a number of specialists, all of whom share the overall goals and research methodology of the group. ... Today, we are still largely rooted at the level of template matching based on parametric representations of the speech signal, with no account taken of the linguistic relation to speech.

It is true that the corpus had not been designed to satisfy the particular needs of speech processing research, with respect to technical quality of the recordings, choice of speakers, situations, topics, etc. Still, LLC was the only large corpus of its kind that was freely available, prosodically analyzed, machine-readable, and familiar to us. It had been the aim of the compilers to provide samples of a wide range of spoken English functions: in monologue and dialogue, intimate conversation and public discussion, face-to-face and telephone conversation, spontaneous and prepared speech, and so on (see p 12). Yet special care had been taken to include spoken language in what we consider its typical form, ie spontaneous speech, when the speaker is thinking on his (or her) feet and is unaware of being recorded. Even if this is a type of speech that is far removed from the most obvious uses of synthetic speech (such as announcements over a loudspeaker), it should nevertheless be valuable in providing insights into what can safely be considered the most characteristic form of spoken - as compared with written - discourse.

## 2.5 Pausing and chunking

During the last decade there have been several studies of spoken English and the relation between speech and writing (for example Halliday 1985a and 1985b; Chafe 1979, 1980b, 1987; Chafe & Danielewicz 1987; Biber 1986a, 1986b, and 1988). In his most recent study, which includes an extensive and critical discussion of previous work, Biber (1988:199) maintains that

> there is no single, absolute difference between speech and writing in English; rather there are several dimensions of variation, and particular types of speech and writing are more or less similar with respect to each dimension.

Yet, there are certain differences that must be considered absolute, and hence important to text-to-speech conversion. One feature that is exclusive to speech is the variation between spoken and silent stretches in the time dimension and the length of the silent stretches. A large part of our speaking time - sometimes more than half - is occupied by pauses, and 'the intermittent silences between chunks of speech, i.e. continuous vocal sequences sandwiched between two pauses ... have been shown to be a characteristic phenomenon in speech production' (Goldman-Eisler 1980:143). This feature of segmentation and pausing, which has been found to be a valuable commodity for interpreters, is bound to be important also for the general language user, as stated by Déjean Le Féal (1982:237):

> If we accept the hypothesis that the division of speech-flow into short segments, acoustic relief and both accidental and deliberate redundancy all have as their origin the concomitance of the processes of ideation and expression in the impromptu speaker, then we must conclude that it is the presence of all these characteristics at one time (a necessary consequence of their common origin) that helps the listener to understand the speaker's meaning.

While it is obvious that pauses constitute an important part of spoken discourse (see further Chapter 8), there are diverse views about the reason for pausing. What seems particularly relevant to work towards improved synthetic speech is the view that 'pauses may serve not only to make time available for the speaker's cognitive processes, but also to assist the listener in his task of understanding the speaker' (Butterworth 1980:157).

For an illustration of pausing in our corpus, compare two different representations of the beginning of text S.12.6, which is an *LLC* monologue that will be extensively used in this book. Figure 2:1a shows the talk transcribed in ordinary orthography and without punctuation. The text in Figure 2:1b is identical, but it is now displayed in segments (one per line) according to how the speaker paused (short pauses are indicated by periods and longer pauses by dashes, see p 7).

**Figure 2:1a.** A spoken text without prosodic segmentation. (Beginning of text S.12.6)

well rather than give a talk about the history of Stoke Poges I felt it might be a little more interesting to you all to hear about my own life lived and growing up in this wonderful village of Stoke Poges I attended Stoke School and I must say I was taught very thoroughly the three Rs funnily enough my father went to the same school and he was one of the first pupils before that he used to go to the school next door to here and pay a penny a week along with all the other village boys for his education considering his schooling must have stopped at about fourteen years his beautiful copperplate writing and his reading with understanding was really remarkable

**Figure 2:1b.** A spoken text with pause-defined segmentation.

well rather than give a talk about the history of Stoke Poges •
I felt it might be a little more interesting to you all –
to hear about my own life –
lived and growing up in this •
wonderful village of Stoke Poges – – •
I attended Stoke School •
and I must say I was taught very thoroughly the three Rs –
funnily enough my father •
went to the same school •
and he was one of the first pupils –
before that •
he used to go to the school next door to here •
and pay •
a penny a week along with all the other village boys •
for his education –
considering •
his schooling must have stopped •
at about fourteen years •
his –
beautiful copperplate writing •
and his reading •
with understanding was really remarkable – –

For any attempt at producing natural-sounding synthetic speech from the text in Figure 2:1a it would certainly be useful to know where pauses might occur. Even if we cannot now give rational explanations for the placing and function of all the pauses, they clearly do not occur randomly. The graphic structure of spoken language with its intermittent bursts of brief chunks of vocalization surrounded by stretches of non-vocalization presents a radically different picture from that of written language with its closely-knit strings of sentences bounded by periods or equivalent punctuation marks. However, the transfer of language material from writing to speaking is no straightforward matter: there is neither a conventional system of pausing in speech corresponding to punctuation in writing, nor is there a one-to-one relation between punctuation and intonation.

Whereas the sentence of writing is well documented in linguistic theory, grammatical description and teaching materials, there are comparatively few statements about a corresponding unit of speech - there is in fact no agreement about how the unit should be named, let alone defined. The sentence is not an obvious candidate here. There is little in the phonetic signal that provides evidence of spoken sentences, and it is difficult, problematic and perhaps irrelevant (see Crystal 1980, Chafe & Danielewicz 1987) to analyze speech in terms of sentences. The segmentation of speech may be viewed as different types of unit (cf Altenberg 1987:47):

**A prosodic unit** manifested as a coherent intonation contour optionally bounded by a pause and containing a salient pitch movement with a principal accent ('nucleus', 'tonic', 'main stress', etc) normally occurring at the end of the unit.

**A cognitive unit** maximally consisting of one newly activated concept - which the speaker brings into his focus of consciousness - and (optionally) some already active (or semi-active) concepts.

**A textual unit** consisting of a part typically carrying new information and optionally preceded or (less commonly) followed by a part carrying given information.

**A grammatical unit** containing at least one phrase or clause element but often a more extensive grammatical structure.

**A physiological unit** 'organized about vegetative constraints of respiration and the physiology of the lungs and larynx' (Lieberman 1986:240).

Depending on where the focus is laid, such segments are variously called 'tone units', 'tone groups, 'intonation units', 'information units', 'idea units', 'chunks', 'phonemic clauses', 'breath groups', and so on.

Since the basic prosodic unit used in the transcription of LLC is the tone unit (see Svartvik & Quirk 1980:14), it was natural to adopt this term for the TESS project. In our approach, the tone unit is contour-defined, not pause-defined: it is only optionally bounded by a pause, and there is no one-to-one relationship between tone units and pause-defined segments. Figure 2:1c shows the same text fragment as in the two previous figures, but now with the text segmented into tone units as used by the speaker in the recorded text (tone units are here printed one per line, and nuclei indicated by capital letters). A comparison between these two segmentation types shows that the 125 words of Figure 2:1a are divided up into 22 pause-defined segments (Figure 2:1b) but into 29 tone units (Figure 2:1c) - averaging, respectively, 5.7 and 4.3 words.

## 2.6 The tone unit in speech production and reception

While the system of transcription made the tone unit the natural unit for us to work with, the decision to place this prosodic unit centre-stage was not simply a matter of linguistic heritage. In fact, one of the main reasons for starting the TESS project was the observation that there seemed to be no widespread recognition of, let alone attempt at introducing, such a unit of segmentation in work on speech synthesis and speech recognition.

There is considerable support for the opinion that the division of stretches of languages into tone units is not arbitrary but has linguistic significance. Segmentation into tone units (or some similar form that is motivated on prosodic, cognitive, textual etc grounds) appears to be a basic requirement of natural and intelligible speech. The primary reason why the sentence is not the relevant unit of speaking is that the sentence is either a grammatical or graphemic construct (depending on how it is defined). The spoken unit has to be appropriate for the conditions under which spontaneous speech is produced, in particular the limited possibilities of planning. One striking difference between the two types of unit is length measured in number of words: the average sentence in written English is much longer than the tone unit in spoken English. The average number of words per sentence is 19 in LOB and 18 in Brown (with peak figures of, respectively, 28 and 24 words per sentence in government documents and similar texts; see Johansson & Hofland 1989, vol 1:17). By contrast, in a study of spoken English the average tone unit length was found to be 4.5 words, with a textual variation ranging from 3.9 words to

**Figure 2:1c.** A spoken text with segmentation into tone units (words with nuclear syllables are in capital letters).

well rather than give a TALK about
the history of Stoke POGES
I felt it might be a little more INTERESTING to you ALL
to hear about my own LIFE
lived and growing UP
in this wonderful village of Stoke POGES
I ATTENDED
Stoke SCHOOL
and I must SAY
I was TAUGHT
very THOROUGHLY
the three Rs
funnily enough my FATHER
went to the same SCHOOL
and he was one of the first PUPILS
before THAT
he used to GO
to the school next DOOR to here
and PAY
a PENNY a WEEK
along with all the OTHER village BOYS
for his EDUCATION
CONSIDERING
his schooling must have STOPPED
at about fourteen YEARS
his beautiful copperplate WRITING
and his READING
with UNDERSTANDING
was really REMARKABLE

5 words; the speech rate was 1.9 seconds per tone unit, with a textual variation ranging from 1.1 seconds to 2.5 seconds (Altenberg 1987a:22f).

The written sentence is clearly too long a unit to handle in speech, particularly in interactive discourse, where planning and execution partly overlap. The real-time processing in speech requires a short prosodic segment, such as the tone unit, for comprehension as well as production. As noted by Boomer & Laver (1968:9), there is good evidence for believing that 'the tone-group is handled in the central nervous system as a unitary behavioural act, and the neural correlates of the separate elements are assembled ... before the performance of the utterance begins'. Similarly, Welin (1979:45) states:

> It is not necessary ... to assume that these smaller phrases are ever combined into larger syntactic units; they might very well be semantically interpreted immediately, and the pointers between them could have an entirely semantic character. ... We seem to attend to syntactic form only in rather short chunks of speech when listening.

To Chafe idea units are 'linguistic expressions of focuses of consciousness' (1980b:15); to Halliday the tone group represents 'a meaningful segment of discourse. Each tone group is, so to speak, one quantum of the message, the way the speaker is organising it as he goes along' (1985a:53). We are aware that speech perception is an active rather than a passive process and 'active hypothesizing concerning the intended message is clearly a part of the speech perception process' (Bond & Garnes 1980:128). Speech perception employs heuristic strategies such as paying attention to stress and intonation patterns, and identifying a suitable prosodic chunk by applying one's grammatical knowledge of the language. We may not be aware of such mental processes when we use our mother tongue, but they become crystal clear when we have to take an active part in a conversation in a foreign language in which we are not proficient. The ability to make 'intelligent guesses' is an important part of our native speaker heritage.

It is a familiar fact that 'the most difficult of the speech processing tasks is the understanding of continuous spoken utterances from unknown speakers using fluent English syntax' (Fallside & Woods 1985:xviii). Such problems in automatic processing are not immediately obvious to human processors who live in societies where it is a natural part of education and everyday adult life to be exposed to the written sentence (with institutionalized punctuation) which is made up of words (surrounded by spaces or some other boundary indication) which consist of certain combinations of letters (limited in number and neatly ordered, without overlap, one after the other). The problems facing the recognition of natural speech are drastically different, which explains why this process is difficult to achieve by mechanical means:

> In fluent speech there are no physical cues that consistently and reliably indicate where one word ends and the next begins. ... At a conservative estimate, less than 40% of all word boundaries are marked by some physical event, and most of these coincide with a constituent boundary.
>
> The impression that words in fluent speech are separated in time is an illusion. Words exist in the mind of the perceiver - not in the stimulus. (Cole & Jakimik 1980:138).

It is clear, then, that we cannot rely on phonic substance alone for recognizing speech. In order to program the machine properly, we must know more about how man processes speech. Obviously, we use all the linguistic levels in decoding a message. Assuming that a large part of our competence in understanding a spoken message consists of intelligent guesses, those guesses are based on phonic substance, on our knowledge of vocabulary items and of the likelihood of specific collocations of words, on our knowledge of grammatical forms and the possibility of grammatical constructions occurring with specific grammatical items, on our knowledge of our interactant's personal linguistic habits, etc. A poor recording, such as a surreptitious one is bound to be, probably reflects quite truthfully the extent to which bits and pieces of the stream of discourse disappear in the transmission from sender to receiver, yet native speakers can decode the message thanks to their remarkable ability of linguistic reconstruction. At the present state of our knowledge we can only guess how this is possible; it is likely to take place on several different levels. In one suggestion as to where to look for the thread of discourse, spontaneous speech has been compared to jazz improvisation where one note gives the next according to the basic chord sequence (Householder 1982).

Assuming that the tone unit is a natural segmentation unit for both production and comprehension, it would be of value for work in speech synthesis and recognition to have a better knowledge of the linguistic structure of the tone unit. As can be seen even in a small fragment like that in Figure 2:1c, the chunking into tone units is not a random process; hence we can assume that native speakers have some kind of built-in 'tone unit grammar'. This mental grammar allows them to plan, execute, and retrieve a message in chunks which do not have a constant one-to-one relationship with any one grammatical unit: tone units may correlate with the whole range of grammatical units, as can be illustrated even by our small corpus fragment, many of which are grammatical phrases, as in:

| | |
|---|---|
| Noun phrase: | Stoke SCHOOL |
| Verb phrase: | lived and growing UP |
| Adverb phrase: | very THOROUGHLY |
| Prepositional phrase: | in this wonderful village of Stoke POGES |

There are also clauses, such as the nonfinite infinitive clause

to hear about my own LIFE

but the grammatical units of sentences and clauses are not regularly coextensive with tone units; instead we find other combinations of sentence and clause fragments, as in the last part of the illustration:

CONSIDERING
his schooling must have STOPPED
at about fourteen YEARS
his beautiful copperplate WRITING
and his READING
with UNDERSTANDING
was really REMARKABLE

So far the corpus illustration has only included tone units, nuclei and pauses. A complete transcription of course provides much more information about the actual speech event, including a number of other prosodic features (onset, boosters, stresses, tones, etc) as can be seen in Figure 2:1d.

## 2.7 Research procedures and results

The TESS research project involved three major tasks, which are documented in this book and other publications referred to in the following chapters.

The basic task was the construction of an automatic tagging and parsing system, which is described in Chapters 3 and 4. Needless to say, the chapter order in this book does not reflect the order of the research procedures. An example of this is the tagset, which did not come out of the blue but was the end-product of a series of studies and tests carried out over a long period of time and in parallel with other tasks. However, the order in which the two

**Figure 2:1d.** A spoken text with complete LLC transcription. (For a key to symbols, see p 7.)

well ||rather than 'give a △TALK a'bout■
the ||history of Stoke △POGES■ •
I ||felt it 'might be a △little more △INTERESTING {to you ||ALL■ }■ –
to ||hear a'bout my △own LIFE■ –
||lived and △growing UP■
in ||this • 'wonderful △village of Stoke △POGES■ – – •
I AT||TENDED■
||Stoke SCHOOL■ •
and I ||must SAY■
I was ||TAUGHT■
||very THOROUGHLY■
the ||three RS■ –
||funnily e'nough my △FATHER■ •
||went to the 'same SCHOOL■ •
and ||he was 'one of the 'first △PUPILS■ –
be||fore THAT■ •
he ||used to GO■
to the ||school next △DOOR to 'here■ •
and ||PAY■ •
a ||{PENNY} a WEEK■
a||long with △all the OTHER 'village BOYS■ •
||for his 'EDU△CATION■ –
CON||SIDERING■ •
his ||schooling 'must have STOPPED■ •
||at about △fourteen △YEARS■ •
||his – △beautiful △copperplate WRITING■ •
||and his "△READING■ •
with ||UNDERSTANDING■
was ||really RE△MARKABLE■ – – –

tagsets are presented (in Tables 3:1-2, pp 96ff) does reflect the research process in that the new tagset grew out of the old: it has finer distinctions and totals over 200 word-class tags. Compared with the number of traditional word classes, this figure may seem like grammatical overkill, yet it can safely be predicted that this is only the beginning of a general development in the direction of greater 'delicacy' in word classification. There are at least two reasons for this development: one is the increased demands on a classification that is to be used for automatic (as opposed to human, pedagogic) processing of language texts, another is the very use itself of the computer as a linguistic research tool in that this tool permits refinements in a classification that are not practical or possible to handle with antediluvian, manual methods. With an error rate of between 3 and 6 per cent, our tagger must be judged efficient, and able to give other competing taggers a match, especially in view of the abundance of tag categories in our system. As for the design of the parser, it was time more than anything else that prevented us from carrying out all our ideas, so that we now have only a very simple parser. At the same time - and this may sound like an excuse - it is interesting to see how much can be achieved with a simple parser. While still experimental and tested only on a limited amount of text, the parser does a good job of splitting up written text into grammatical phrases.

A second task was to improve our general knowledge of lexical, prosodic, and grammatical features in spoken English. Such aspects are discussed in a monograph (Altenberg 1987a) and several papers, some of which are included in this volume. Two chapters are specifically devoted to lexical aspects. Chapter 5 presents an analytical model of the structure of spoken interaction in terms of four hierarchical levels (exchange, turn, move, and act) and presents results from a study of lexical items that are peculiar to - and usually also very frequent in - spoken English. A comparison with traditional word classes shows that spoken discourse items (*well, now, you know, you see, I mean*, etc) are more frequent than prepositions, adverbs, conjunctions etc. Differences in usage are noted between different speech situations (monologue and dialogue, face-to-face and telephone). The discussion of lexical items with pragmatic functions that are difficult to describe in traditional terms is taken up again in Chapter 6, which deals with inadequacies in the treatment in dictionaries of speech-specific phenomena: the use of intonation to differentiate adverbial functions and the use of spoken discourse items.

The next four chapters are concerned with aspects of prosody. Chapter 7 discusses various prosodic features within the tone unit, in particular the booster (ie a step-up in pitch in the intonation contour), but also other pitch-prominent features such as the onset and the nucleus. The booster occurs frequently: in the present material every second tone unit contains such a step-up. Although it has not been possible to include all pitch-prominent features in the present set of rules - one problem with the booster is that it has low predictability - such phenomena will clearly have to be considered in a more advanced system of text-to-speech conversion. A striking shortcoming of synthetic speech is its monotonous and lifeless character, and it is therefore essential to try to introduce a more human-like pitch pattern in order to achieve both better comprehension and greater tolerance.

The abundance of pauses has already been mentioned as a basic characteristic of speech. The occurrence of pausing in LLC texts is discussed at some length in Chapter 8, including not only pauses 'proper', ie silent pauses, but also filled or voiced pauses like *ə:m* and verbal fillers like *I mean, you know,* and *I see* (many of which have already been encountered in Chapter 5 - where however they travelled under cover as 'discourse items'). This chapter provides data on pausing differences in monologue as compared with dialogue and also on different functions, combinations and positions of the three types of pause. As for position, the majority of the pauses occur between clause elements - and not between clauses, as has been stated in some earlier reports on research. Such differences need not be contradictory, however, since they can be explained as due to normal variation to be expected among different speakers and speech situations.

For a program that aims at converting a written message into a spoken message it is obviously important to make use of whatever correlative ties there are between the two channels. The question which Chapter 9 tries to tackle is how reliable 'adverbial punctuation' is as a cue for prosodic breaks. The results are inconclusive, which is only to be expected of such a heterogeneous grammatical category as the adverbial with its wide range of realizations, positions and functions. Certain uses of punctuation do provide cues, but punctuation practice alone cannot be taken as a criterion of prosodic separation.

Chapter 10 is a note on an experiment designed to combine synthetic voice output with graphic representation on screen or paper, in addition to providing some basic statistical information. It was a small project as part of a large thought: to be able to submit to the computer, in machine-readable form, a written text which, automatically and in turn, is grammatically tagged and parsed, prosodically segmented, vocally synthesized, and graphically reproduced.

The third task of the TESS project was setting up rules for segmentation of written text into tone units. By drawing on results from previous research and addressing here the central task of text-to-speech conversion, the topic covered in the last three chapters represents the culmination of the whole project Text Segmentation for Speech. In Chapter 11 it is shown that it is possible to achieve good, but of course not infallible, prediction by combining certain grammatical information (such as type of clause structure) in a 'Matrix Rule' with a 'length factor' (where length is measured in number of words). The success rate is over 90 per cent for predicting prosodic separation of clauses, but lower for predicting breaks between phrases and clause elements, which are better defined in text-linguistic and discourse-functional rather than grammatical terms.

In Chapter 12 the segmentation rules are applied to a written text, an American newspaper editorial in the Brown Corpus. A large number of rules are presented and their success and failure rates discussed. The Prolog computer program is documented in Chapter 13. For a preview of what it can achieve, here is an illustration: first in (1) a fragment of the editorial as it originally appeared when printed in the newspaper, and then in (2) the same fragment as it came out after automatic processing by the segmentation program (see further pp 320ff).

(1) He had been involved in countless schemes to do away with democratic leaders in neighboring countries such as President Romulo Betancourt of Venezuela. It was a sort of poetic justice that at the time of his own demise a new plot to overthrow the Venezuelan government, reportedly involving the use of Dominican arms by former Venezuelan Dictator Marcos Perez Jimenez, has been uncovered and quashed.

(2) He had been involved in countless schemes■
to do away with democratic leaders in neighboring countries■
such as President Romulo Betancourt of Venezuela■
It was a sort of poetic justice■
that at the time of his own demise■
a new plot to overthrow the Venezuelan government■
reportedly involving the use of Dominican arms■
by former Venezuelan Dictator Marcos Perez Jimenez■
has been uncovered and quashed■

With a success rate of over 90 per cent, the segmentation program must be said to be highly satisfactory. Yet much remains to be done in order to achieve a complete text-to-speech conversion package with the inclusion of many more prosodic features than tone unit chunking.

## 2.8 Postscript

This chapter does not aim at presenting a summary of the results of the research presented in this book, let alone the whole project. Yet, at the end of the project and at the beginning of a new decade, I want to conclude the chapter with some very general comments based on our experience from this field of research, and also to speculate about its future. Some of the views have support in our research, but others have no such foundation and are unashamedly subjective; however, they are both offered here for what they are worth.

### Corpus

Since so much of our work at Lund has focussed on the compilation, computerization and use of text corpora, we may unintentionally have given the impression that the corpus-based approach is our sole favourite. If so, let it be said that I for one do not want to exaggerate this form of data to the exclusion of others. At the same time, 'computational linguistics' should not be equated with artificial intelligence. This is a plea for redressing the balance between the two approaches, both of which are clearly needed to solve the problems at hand. On the TESS project we have not been able to make as extensive use of the available corpus as we had hoped. In future research linguists are likely to work with numerous, very large, and different corpora providing empirical data from a wide range of language uses. Corpus-based natural language processing will have many practical applications but also have a greater impact on linguistic theory than in the past.

### New techniques

From various quarters we have noted a growing interest in the 'real data' that a corpus can offer. One of the reasons for the current popularity of corpora is technical: the personal computer - the 'Mighty Micro' - has changed the working conditions of the linguist in providing convenient access to large samples of real language but also new and efficient ways of handling empirical language data. It will be a safe prediction that, so far, we have only seen the modest beginnings of this development: in the nineties, which will bring improved storage capacity (for example CD-ROM), more adequate software (such as programs for organizing and retrieving large amounts of linguistic data), widespread access to the new techniques (audio recordings on compact disc, access to data by electronic mail, international computer networks, etc), there will be even more efficient ways of working with extensive empirical data.

## Tagging and parsing

Apart from being a necessary part of text-to-speech processing, the automatic analysis of texts by means of tagging and parsing offers the linguist an attractive intellectual challenge; also, it provides insights into the properties of language (grammatical, lexical, pragmatic, prosodic, etc) that are not easily available from a manual mode of analysis. With its capacity for speedy performance and tireless repetition the computer makes a superb tool in, for example, testing out new ideas and experimenting with different classifications. As a result, in recent computer-based research we have witnessed a development towards a dramatic increase in the number of lexical categories with an emphasis on the hierarchical relation of the classes of words. A success rate of over 90 per cent for automatic word class assignment, working with over 200 categories, shows that tagging is indeed a realistic proposition; parsing, ie automatic syntactic analysis, is of course a more demanding task but good progress is predictable in view of the worldwide effort that is going into this type of research.

## Chunking

For achieving better results in automatic text-to-speech processing it is necessary to perform a segmentation of the input text into some type of unit - whether it be tone, intonation or information unit - that is more appropriate than the written sentence to *human* speech processing. It is a fact (which seems to have been largely neglected both in work on speech processing and language teaching) that prosodic chunking and stress-timed rhythm are vital factors in perception and production, certainly as far as English is concerned. For this project it was necessary to restrict the aim by concentrating our research on chunking, ie segmentation. This search has been quite successful. As shown here, reasonable chunking can in fact be performed automatically. Yet it is still an open question how far it is possible to achieve a complete text-to-speech conversion package which includes a more extensive set of prosodic features than tone unit chunking. The occurrence of other, more semantically related, features such as boosters, tonicity and tonality are likely to be more difficult to predict - unless you are a mind-reader.

## Human speech processing

The insights into the principles of automatic prosodic segmentation are also valuable for a better understanding of the psycholinguistic processes involved in speaking and listening. My own view of how this takes place, which draws on the ideas and research of many other people, runs along the following lines.

In human speech, the units for production and recognition are limited in size to prosodic segments that can be contained in short term memory, which means chunks averaging around five words - hence considerably shorter than the average written sentence (which, on average, is more than three times as long). These chunks are processed, more or less one by one, without the speaker/listener necessarily operating with the larger linguistic unit of the sentence. In these chunks, which are frequently - but not necessarily - enhanced by pause boundaries, some words are more important than others, as indicated by accenting features, in particular the nuclear tone with its salient pitch movement. These accented parts may well play a central role in human speech processing, such that the nucleus marks the end of the focal element in the speaker's/listener's mind as new ideas are successively activated. It is likely that human processing takes place, not in a simple word-by-word fashion and straightforward left-to-right order, but by various (as yet little-known) operations such as caterpillar-like increments, shunting and backtracking. The need for real-time processing in the spoken mode make very special demands on the participants' ability of speedy construction, retrieval, interpretation - and guessing. For successful communication, therefore, the human mind is bound to use a combination of creative syntactic construction and retrieval of prefab lexical items and more or less frozen collocations. To try to state how this takes place should be one of the great challenging tasks for linguists in the nineties.

## New bearings

The orientation of much linguistic research is undergoing change. The TESS project may be seen as one instance of this, in that we entered an arena that presented new problems and challenges. In retrospect, I think this is a good development for humanistic subjects: it calls for more academic cross-fertilization and fresh approaches to old problems which, hopefully, will lead to a better understanding of the complexities of natural language and the marvel of human language processing. There is, in this field, a real need for people who have experience from working with 'real' language data:

> Much remains to be done before text-to-speech synthesis systems reach the stage where they can be mistaken for human speakers and thus pass the 'Turing-test'. But interestingly, at least for the time being the main limiting factor for improvement of the quality of text-to-speech conversion is not hardware or digital audio synthesis technology. Major advances are dependent on the availability of explicit linguistic rules for allophonic variation and prosody and efficient implementations of these in synthesis devices. (Halvorsen 1988:214)

Thus, success in the attempts at improving the quality of text-to-speech conversion is more dependent on linguistics than technology. Will the nineties be the decade for Renaissance men and women in speech processing?

## Note

Some contemporary men and women were kind enough to read an earlier version of this chapter. I am indebted to Bengt Altenberg, Göran Kjellmer, Muriel Larsson and Gunnel Tottie for their helpful comments and suggestions. I bear sole responsibility for anything the reader may take exception to.

# 3

# Tagging and parsing on the TESS project

*Jan Svartvik*

## 3.1 Why tag and parse?

As described in Chapter 1, the London-Lund Corpus of Spoken English is available in the form of an orthographic word-by-word transcription with an accompanying prosodic analysis. For many purposes, however, these two 'levels' - called 'text' and 'prosody' - offer only limited possibilities for the use of the corpus. In a written corpus, where they may be said to correspond to 'text' and 'punctuation', respectively, the problems are similar but not identical, one reason being that punctuation is institutionalised and its properties generally far better known than those of prosody. Some examples will help to clarify what the limitations may be.

It is a straightforward enough task to search for instances of a word (ie letters with spaces on either side) like *as*. For a subset of the corpus, such as a single text, it can be managed even with the aid of a standard word processing package, although it is of course preferable to use more dedicated text retrieval software tools.[1] However, the task is far from straightforward when it comes to restricting the search to, say, the conjunction *as* (*As I was just saying ...*) as distinct from the adverb (*John was quick, but I was just as quick*) or the preposition (*Speaking as your leader I want to say that ...*). Yet, with a highly frequent item like *as* - there are over 15,000 such tokens in Brown and LOB - we may not be content with isolating all instances of the conjunction *as* but want to narrow the search even further to include its use only in specific clause types, such as temporal (*As they were about to enter ...*), reason (*As she had run out of cash ...*), concessive (*Naked as I was ...*), similarity (*He cooks as his mother did*), or proportional (*As the lane got narrower, it became more difficult to keep the bus on the road*).

Unlike this first example with *as*, a single word with different functions, there are other types, such as different words with certain common morphological attributes like suffixes. If, for example, we want to retrieve all instances of adverbs ending in *-ly*, there is no simple way of restricting such a search to just adverbs. With the net cast so wide, the catch will include not only *rapidly* and *irretrievably* but also *only*, *supply*, and *fly*.

As a third example we may choose the grammatical category of 'adverbial', which features a wide range of semantic roles (space, time, process, modality, degree, etc) and realization types (see Quirk et al 1985:489), for example:

| | |
|---|---|
| Adverb phrase: | She called *early.* |
| Noun phrase: | She called *last week.* |
| Prepositional phrase: | She called *in the evening.* |
| Verbless clause: | She called *though obviously ill.* |
| Nonfinite clause: | She called *while waiting for a plane.* |
| | She called *hoping for a job.* |
| | She called *to ask for an interview.* |
| Finite clause: | She called *after she had read the advertisement.* |

With a 'raw' form of input consisting of an orthographic-cum-prosodic transcription, there is of course no way to ask for an abstract grammatical category like 'adverbial' and hope to get a comprehensive list with the structures that function as the clause element 'A'.

To make successful searches of the types just illustrated it is necessary to add other levels to the input, ie mark the words of the text with tags - hence the term 'tagging'. By tagging, then, we mean assigning, preferably automatically, a lexical-grammatical description (word-class tags) to the words in a text; or, in the words of Francis: 'supplying each word in the Corpus with a symbol indicating its place in a taxonomy based on surface syntactic function' (1980:198).

We reserve the term 'parsing' for marking grammatical levels that are higher than that of word class so that, in our system, tagging (for example assigning the word class tag 'adverb' to *early*) precedes parsing (assigning the clause element label 'adverbial' to *while waiting for a plane* - and, of course, also to *early* - in the examples above).

## 3.2 Tagging spoken and written English

On the TESS project we have done some tagging of both spoken and written English. Unlike Brown and LOB, LLC has not yet been tagged in its entirety - at least not far as we know. We tagged ten LLC texts. The main reason why LLC has not been completely tagged by us is simply that our resources have not, so far, allowed this type of extensive tagging; for the end-product of the TESS project, which was text-to-speech conversion, the tagging had to be done on written, not spoken texts. Furthermore, it was not clear to us what type of tagging is the most suitable for spoken language. Finally, even if it was possible to agree on some criteria for *one* system of tagging, it would still not be adequate for all types of corpus use: different tasks require rather different taggings. To avoid misunderstanding, this is not to say that we are against a comprehensive tagging of LLC - it is indeed a suggestion we are happy to pass on to somebody in quest of a future project and in possession of adequate energy and funding. Ideally, use should then be made of a tagset which is appropriate to the spoken material but also compatible with the Brown/LOB tagset (see Johansson & Hofland1989).

In computational linguistics using the corpus-based approach there is a heavy reliance on large amounts of authentic data as a basis for making probabilistic statements (see, for example, Leech 1987). In his controversy with the 'majoritarian' trend in computational linguistics, ie the artificial intelligence paradigm using the non-probabilistic approach, Sampson quotes sentences which are not authentic but 'toy subsets of artificially simple linguistic forms' (1987:17), such as the following:

> Whatever is linguistic is interesting.
> A ticket was bought by every man.
> The man with the telescope and the umbrella kicked the ball.

For contrast, he quotes some other sentences which are samples of authentic language, drawn from the LOB Corpus with the use of random techniques, such as the following:

> Sing slightly flat.
> Such events and such experiences merit a special designation.
> Say the power-drill makers, 75 per cent of major breakdowns can be traced to neglect of the carbon-brush gear.

Our LLC data are of course far removed from any written data, whether authentic or not, and a large part of our spoken material easily qualifies as 'messy' and 'ill-formed'. A corpus with interactive speech like LLC presents a very special challenge to the linguist in that there are no sentences or clauses to be easily found, ie not even the basic grammatical units are set out in the text. Here is a random example, first given in a straightforward orthographic transcription without prosody (but with speaker shift indicated):

(Text S.3.2.196-216)

| | |
|---|---|
| B | no no Lyle just sort of gave me a sort of you know just said one thing and I said «I must see you 4 to 5 sylls» (laughs) |
| A | yeah yeah yeah ə: is to ferry up the M one to wherever you're living |
| B | m |
| A | with this man |
| B | which we don't know yet (laughs) |
| A | which you don't know |
| B | m |
| A | ə:m ei pair of tape-recorders ə:m and ə: that you work at home |
| B | m |

Including prosody gives a clearer structure and makes the text easier to understand. See the following extract, which is the same passage supplied with prosodic analysis (from Svartvik & Quirk 1980:786).

| | |
|---|---|
| **B** | 196 ☆ǁno NǑ■ 197 ǁLyle 'just sort of☆ ▵GÀVE me■ 198 a ǁsort of ǁYǑU 'know■ 199 ǁjust 'said ☆▵ǑNE☆ ▵{ǁTHÌNG■}■ 200 and ǁI said «I must 'see YǑU■ 201 ǁ4 to 5 sylls»■ 202 (laughs · ) |
| **A** | 203 ☆ǁYÈAH■☆ 204 ǁYÈAH■ 205 ǁYÈAH■ · 206 [ə:] · ǁis to · '∆ferry 'up the ▵M · ∆ǑNE■ · 207 ǁto wher▵ever you are '∆LÌVING |
| **B** | 208 ǁ[m̄]■ |
| **>A** | 207 'with this MĀN■ · |
| **B** | 209 ǁwhich we don't KNÓW■ · 210 ǁYÉT■ 211 ☆(laughs · )☆ |
| **>A** | 212 ☆ǁwhich☆ you 'don't KNÓW■ |
| **B** | 213 ǁ[m̄]■ · |
| **A** | 214 [ə:m] · ǁ[ei] ∆pair of ∆TÀPE-RECORDERS■ – 215 [ə:m] ǁand [ə:] ∆that you ∆work at ∆HǑME■ |
| **B** | 216 ǁ[ḿ]■ |

However, sentences are still not easily located and authentic, informal, interactive talk presents problems even at the word level. Yet issues relating to this level are commonly (but erroneously) assumed to be simple and to have been settled by our forefathers several hundred years ago, such as the number of word classes required for the description of a language. One thing an authentic corpus helps to drive home is the weakness of the traditional word-class system, as will be further discussed in Section 3.4.

## 3.3 Levels of analysis

In our approach to tagging and parsing we recognize several linguistic 'levels':

TEXT, ie the orthographic representation of the spoken material (as in the first example above);
PROSODY, ie the prosodic analysis of the audiotapes (as in the second example above);
WORD CLASS, where each word is assigned a word class tag;
PHRASE, where the texts are analysed in terms of five types of grammatical phrase: adverb phrase, adjective phrase, verb phrase, noun phrase, and prepositional phrase (Quirk et al 1985:38ff);
CLAUSE, with an analysis into the clause elements Subject, Verb, Complement and Adverbial (Quirk et al 1985:49ff);
DISCOURSE, to deal with those features which are characteristic of spoken English but not easily or adequately accounted for in terms of 'orthodox' grammatical analysis (see Chapter 5).

The first two, text and prosody, are directly available in the transcriptions of the corpus, whereas word class, phrase, clause and discourse are not.

A basic element in our approach has been the grammatical analysis of tone units. Our approach is similar to that of the Lancaster project in that we stress the importance of real speech: 'It is only by closely studying a corpus of natural data that one can begin to understand the problems, let alone find algorithms to solve them' (Knowles & Lawrence 1987:148). In view of our emphasis on the importance of the tone unit rather than the sentence in speech processing, we also agree with the gist of the statement 'the assumption that one needs to parse a sentence before doing any interesting linguistic processing is almost certainly false' (148). However, the segmentation into tone units may be correlated with several linguistic 'levels', as the term has been used above, and we cannot now say that only one level, such as word class, is adequate for the task. Hence we do not exclude parsing as a useful tool for gaining a better understanding of the principles of human speech processing, as long as 'parsing' need not be understood as a grammatical analysis of whole sentences. In many cases the grammatical phrase is obviously an important element in tone unit formation, and it is therefore a relevant task to identify, and assign grammatical structure to, such cohesive units in the flow of natural speech.

## 3.4 The old and new tagsets

A typical feature of our approach to word-class tagging has been to consider it not an end in itself but rather a stepping-stone to syntactic analysis (parsing); also, we have favoured an open-ended system suitable for different uses. Automatic analysis of genuine texts, spoken or written, requires a close interplay between the different levels of grammatical analysis, and it is therefore necessary to revise, from time to time, the categories on one level in order to achieve a better result on another level. Hence we thought it desirable to design a tagset which was flexible and open for revision during the period of the research project rather than aim at a fixed tagset for general use.

In our first attempts at tagging LLC (see Svartvik et al 1982:51 ff), we had a limited set of word-class tags, hence referred to here as 'the old tagset' (listed in Table 3:1, pp 96ff). In the subsequent TESS project the old tagset was considerably extended in terms of quality (levels of delicacy), and hence also in terms of quantity (number of tags).

Our new tagset (see Table 3:2, pp 101ff) can be said to have three characteristic features. First, it is intended to be a tool for word-class analysis that covers features that are typical of spoken English but not adequately covered in a traditional word class system oriented to written discourse. The insights that we had got from the study of our database, noting in particular the highly frequent use of items characteristic of, or unique to, spoken discourse such as *well, I'm sorry, sort of, you know* called for new tagging categories (see 'D-items' in Chapter 5, pp 137ff).

A second feature is that our tagset comprises finer distinctions than most comparable sets. It was obvious that the traditional word-class categories were too coarse for the parsing program. In particular, the adverb class is so heterogeneous in terms of semantics, lexical realization, and syntactic position that it needed a more delicate subclassification than that used in our old system - and other traditional tagging systems.

A third feature of our tagset is that it has a hierarchical structure. Thus, in the tag <ACres> the first capital letter 'A' stands for the main grammatical category 'adverb', the second capital letter 'C' stands for the functional subclass 'conjunct', and the lower case three-letter combination 'res' stands for the conjunct subclass 'resultive'. The word class 'adverb' is notoriously heterogeneous (see Quirk et al 1985:73) and the customary, simple bifurcation into 'open' and 'closed' class adverbs is not really adequate. Consider, for example, the following two sentences:

> She couldn't see the ship clearly because of the fog.
> Clearly she couldn't see the ship because of the fog.

The two uses of the adverb *clearly* are different in terms of meaning, position and prosody, yet they both qualify for 'open class' membership by virtue of their common *-ly* adverb suffix. In the new tagging system the two exemplified uses are tagged respectively 'adjunct' <AA> and 'disjunct' <AD> (largely following the classification of Chapter 8 in Quirk et al 1985; for a somewhat different breakdown of adverbial classes see a more recent work, Lindquist 1989). The adverb categories are also further subclassified, so that the two instances of *clearly* in the examples are tagged as 'process adjunct' <AApro> and 'content disjunct' <ADcnt>, respectively.

When tagging was first begun at Lund in the late 70s, we did not have access to a system for assigning unified tags to more than one graphic word, which was most unsatisfactory. It is, for example, clearly awkward to have to assign to a fixed phrase like *of course* the two tags 'preposition + noun' (ie the same grammatical description as for a free syntactic construction like *off course*). Fortunately, this could be remedied in the new system. In our approach to the study of spoken language, an important element is the concept of 'cohesive chunks' in the flow of speech and we therefore want to assign a unified tag to them. Such cohesive units may be prosodic ('tone units', see pp 74ff) or lexical, for example complex prepositions like *by means of* and adverb phrases like *on the other hand*. Such lexical chunks are entered in our dictionary and assigned a unified tag; hence the given complex preposition is tagged '*by* <PA1> *means* <PA2> *of* <PA3>'. By including complex prepositions (Quirk et al 1985:669) the category of 'preposition' now totals over 200 types (see Svartvik 1988), eg *according to, together with, in contact with, on top of, with respect to*. Adverb phrases like *on the one hand, on the other hand* are also entered in the dictionary (as contrastive and listing conjuncts), whereas a similar string of words like *on the other foot* is not a dictionary entry but analysed by the grammatical phrase parser as a regular prepositional phrase.

However, in spite of different 'derivational histories', all three phrases will be assigned the clause level status of Adverbial. There are several reasons for the decision to choose this mode of analysis: in particular economy, predictability and prosody.

Other types of cohesive chunks which have been retagged are long strings of determinatives, which now include for example the following:

| | | |
|---|---|---|
| | *the* | diamonds |
| *all* | | diamonds |
| *all* | *the* | diamonds |
| *some of* | *all the* | diamonds. |

In view of the characteristics of spoken English that we wanted to capture, it seemed more important to have a system specifically adapted to our corpus than to conform to the tagging systems already in use for corpora of written English. However, although our system is different from those tagsets, it is not incompatible with the system used for the tagging of Brown and LOB (see Francis & Kučera 1989, Johansson 1986, Johansson & Hofland 1989). While the new tagset was clearly geared to handling the features of spoken English which we felt had been inadequately catered for in previous work on tagging, the tagset is not in any way intended to be used exclusively for spoken English, and we applied it also to written texts (as described in Chapter 4).

The new set of word class categories totals over 200 items, if the tag combinations used to tag contracted 'words' (eg *there's*) are included in the count (see Table 3:2, pp 101ff). Our set is large in comparison with other similar sets: the tagged Brown Corpus uses 179 different wordtags, the LOB tagset comprises 132 tags, and the Leeds tagset 137 tags. In the ongoing work on tagging the LOB Corpus, the general tendency has been to introduce finer distinctions; hence the 'Lancaster tagset' comprises 200 or more members (according to Sampson 1987b).

It must be stressed that our work on tagging and parsing has been very much experimental. We concentrated our efforts on trying out different ways of analysing parts of the corpus rather than on implementing one tagging system to the whole corpus in order to achieve a complete system of text-to-speech conversion (which would clearly be beyond our means anyway). In our experiments with new ways of tagging we have found good reasons to support the view of Crystal:

> It is important not to let the familiarity of the traditional terms obscure these more important issues: words like 'verb', 'adverb', 'grammatical word', and so on, slip

smoothly from the tongue, and for practical economy of reference one has just got to use them and hope for the best. But this should not be allowed to engender a false sense of security: each term has its weaknesses, and its validity must ultimately be assessed in the light of some general linguistic theory. If there is any conclusion at all that would not be premature from this turning-over of largely familiar ground, it is simply that word classes in English are more complex things than is generally supposed; and that before we can produce a set of satisfactory definitions, we need to examine the distribution of single words much more thoroughly. (1967:55)

## 3.5 Parsing at phrase level

'Parsing' is the term we have used for assigning labels to the phrase and clause levels, following word-class tagging. In our parsing experiments, the emphasis has been on the phrase level, and we will therefore restrict the presentation of our approach in the following chapter to this level. Since the phrase level analysis is based on the word-class tagging, reference should be made to the new tagset in Table 3:2. The rules which operate on the word-class tags to produce the grammatical phrases are further discussed and listed in Section 4.3 (pp 110ff); cf also Section 12.3.

The phrase level was analysed in terms of five types of grammatical phrase:

| | |
|---|---|
| Adverb phrases | [APH, see Section 4.3.4] |
| Adjective phrases | [JPH, see Section 4.3.5] |
| Noun phrases | [NPH, see Section 4.3.6] |
| Prepositional phrases | [PPH, see Section 4.3.7] |
| Verb phrases | [VPH, see Section 4.3.8] |

This analysis and the general terminology used for the parsing system are largely based on Quirk et al 1985, to which reference should be made when studying the rules in Chapter 4.

In Chapters 12 and 13 we describe the use of tagging and parsing for segmenting a written input text into appropriate information units.

### Note

1 There are some such packages available, for example 'WordCruncher' (from Electronic Text Corporation, Orem, Utah, USA), 'CLAN' (from Children Language Data Exchange System, Carnegie Mellon University) and 'KAYE, The KWIC Analyser' (from IBM UK Scientific Centre, Winchester, England), but users have to be aware of the possible need for tuning in any tool to the requirements of specific data formats such as prosodic features.

**Table 3:1**

## The old set of word-class tags

| | |
|---|---|
| A | ADVERBS |
| AB | *wh*-adverbs: *where, why, how* ... |
| AC | Closed class adverbs (other than those specified here otherwise), ie those not derived from adjectives: *slow, early, there, just, far* ... |
| AE | Postpositional adverbs: *ago, enough* ... |
| AF | *more* (all uses) |
| AG | *most* (all uses) |
| AI | *very* |
| AM | *much* |
| AN | *not* |
| AP | Adverbs as particles in phrasal verbs (*to set up*) or nouns (*the washing up*), but prepositions in prepositional verbs (*walk up the street*) = PA |
| AR | *no*: 'no more', 'no good' ... |
| AS | *yes, yeah, yea, aye* ... |
| AT | *too* |
| AW | Open class adverbs, ie those ending in *-ly* and derived from adjectives: *easily* ... |
| AX | *there* as existential adverb (locative *there* = AC) |
| AZ | *so* (all uses) |
| +R | comparative: *earlier* = AC+R |
| +T | superlative: *earliest* = AC+T |
| C | CONJUNCTIONS |
| CA | *and* |
| CB | *but* |
| CC | All subordinators except *that* |
| CD | Subordinating *that*: 'She thought that ...' |
| CE | Correlative *as, than* |
| CR | *or* |
| CS | *both, either, neither* |
| E | PREDETERMINERS |
| EC | *all*: 'all the time' |
| ED | *half*: 'half a pint' |
| EE | *both*: 'both days' |
| EF | *double*: 'double the income' |
| EJ | *many*: 'many a day' |

| | |
|---|---|
| EK | *what, such*: 'what bad luck' |
| EL | *quite, rather*: 'quite a girl' |
| | |
| G | RELATIVE PRONOUNS |
| GA | *who, which* |
| GB | *whom* |
| GC | *what* (without a nominal head): 'what he says is true' |
| GD | *that*: 'the car that is parked over there' |
| | |
| I | INTERJECTIONS |
| IA | Interjections 'proper': *oh, gosh* ... Other word-classes used as interjections are coded thus: *fine* JA = I, *God* NP = I, *well* AC = I, *what* RF = I, *thanks* NC+2 = I ... |
| | |
| J | ADJECTIVES |
| JA | Adjectives other than those specified below |
| JB | 'be *able* to', 'be *about* to' |
| JE | Postposed adjective: 'president elect' |
| JM | *few, least, less, little, many, more, most, same, several, such,* + noun: '(a) few (young) people' ...; *own* (all uses): 'my own car' ... |
| JN | Adjectives denoting nationality and/or language: *English* ... |
| JP | Adjectives as noun phrase head: 'the dead', 'the accused', 'the same' ... |
| JQ | Ordinals: *first, second* ... |
| JR | Cardinals: *one, two* ... and pseudocardinals: *fortyish* ... |
| +R | comparative: *finer* = JA+R |
| +T | superlative: 'do one's best' = JP+T |
| | |
| N | NOUNS |
| NC | Nouns other than those specified below |
| NN | Nationality nouns: *Englishman* ... |
| NP | Proper nouns: *England, John* ..., including eg titles before names (*Sir John*) and book titles |
| NX | Nominal abbreviations, acronyms etc: *BBC* |
| +2 | plural: *boy* = NC, *boys* = NC+2 |
| +Z | genitive: *boy's* = NC+Z, *boys'* = NC+2+Z |
| | |
| P | PREPOSITIONS |
| PA | Prepositions other than PD |
| PD | Infinitive marker *to* |
| | |
| R | PRONOUNS OTHER THAN RELATIVE (with non-determiner function) |
| RA | Personal/subjective: *I, he*, ... |

| | | | | |
|---|---|---|---|---|
| RB | Personal/objective: *me, him* ... | | | |
| RC | Personal unmarked for case: *you, it* | | | |
| RD | Demonstrative *that*: 'that's a nice car' | | | |
| RE | Possessive with non-determiner function: *mine* ... | | | |
| RF | Interrogative: *who, whose, which, what* (+ *ever*) | | | |
| RG | Demonstrative *this*: 'this is terrible' | | | |
| RH | Demonstrative *these, those*: 'how much are these?' | | | |
| RJ | *One, ones* (all uses except numeral): 'that's a big one' | | | |
| RM | Pronouns in *-body, -one, -thing* (transcribed as one word): | | | |
| | *somebody* | *someone* | *something* | |
| | *anybody* | *anyone* | *anything* | |
| | *everybody* | *everyone* | *everything* | |
| | *nobody* | *noone* | *nothing* | |
| RN | *else* | | | |
| RP | Quantifiers: | | | |
| | *all* | *either* | *many* | *other(s)* |
| | *another* | *enough* | *more* | *plenty* |
| | *any* | *few* | *most* | *several* |
| | *both* | *less* | *much* | *some* |
| | *each* | *little* | *neither*<br>*none* | *such* |
| RR | Reflexives: *myself, ourselves* ... | | | |

| | |
|---|---|
| T | DETERMINERS |
| TA | *the*, [dhi(:)]: 'the car' ... |
| TB | *my, no, whose, which, whatever*: 'whose car' ... |
| TC | *some, any, enough*: 'any car' ... |
| TD | *his, that*: 'that car' ... |
| TE | *these, those*: 'those cars' ... |
| TF | *a, an*, [ei], [æn]: 'a car' ... |
| TG | *every, each, either, neither, another*: 'every car' ... |
| TH | *much*: 'much trouble' ... |

| V | VERBS<br>MAIN VERBS<br>VA | BE<br>VB<br>(all uses) | DO<br>VD<br>(all uses) | HAVE<br>VH<br>(all uses) | MODALS<br>VM |
|---|---|---|---|---|---|
| +0 | *speak* | *be* | *do* | *have* | |
| | *walk* | | | **'ve* | |
| +3 | *speaks* | *is* | *does* | *has* | |
| | *walks* | **'s* | **'s* | | |
| +D | *spoke* | | *did* | *had* | |
| | *walked* | | | **'d* | |
| +N | *spoken* | *been* | *done* | *had* | |
| | *walked* | | | | |
| +G | *speaking* | *being* | *doing* | *having* | |
| | *walking* | | | | |
| +1 | | *am, *'m* | | | |
| +4 | | *are, *'re* | | | |
| +5 | | *was* | | | |
| +6 | | *were* | | | |
| +8 | | | | | *can, may ...* |
| +9 | | | | | *could, might ...* |

Marginal verbs, eg *need*, are coded thus:

VM+8 She *need* not go

VA+0 She does not *need* to go

Contractions (indicated by asterisk)

| | |
|---|---|
| VB+3*AN | *isn't* |
| VB+4*AN | *aren't* |
| VB+5*AN | *wasn't* |
| VB+6*AN | *weren't* |
| RC*VB+4 | *you're* |
| RA*VB+1 | *I'm* |
| RA*VH+0 | *I've* |
| RA*VB+3 | *he's* (here) |
| RA*VH+3 | *he's* (done it) |
| V%+3 | *'s* when no way of saying whether it is a *be* or *have* form: '*he's* gone' ... |
| RA*VH+D | *she'd* (done it) |
| RA*VM+9 | *she'd* (come if she could) |
| RA*VM+8 | *they'll* |
| V%+7 | *ain't* |

| | |
|---|---|
| X | MISCELLANEOUS |
| XA | Metalanguage: 'the word *like*' |
| XM | /m/, /mhm/ |
| XR | /ə/, /əː/, /əm/, /əːm/ |
| XX | Foreign words: *Kaiser, ad hominem*, formulae: *s log* ..., separate letters: *J O H N* |
| XZ | general 'ragbag' /hm/, /uhuh/ ... |

QUERIES

When in doubt about word-class assignment, either bare '?' or 'NC?', 'VA+D?', etc have been used.

**Table 3:2**

## The new tagset

| TAG | CAT | SUBCAT | SUBSUB or ITEM | EXAMPLE |
|---|---|---|---|---|
| AApro | adverb | adjunct | process | *correctly* |
| AAspa | adverb | adjunct | space | *outdoors* |
| AAtim | adverb | adjunct | time | *often* |
| ABdeg | adverb | wh-type | degree | *how* |
| ABman | adverb | wh-type | manner | *how* |
| ABrea | adverb | wh-type | reason | *why* |
| ABspa | adverb | wh-type | space | *where* |
| ABspa*VB+3 | adverb | wh-type+*s*-form | space | *where's* |
| ABtim | adverb | wh-type | time | *when* |
| ACapp | adverb | conjunct | appositional | *for example* |
| ACcon | adverb | conjunct | contrastive | *by contrast* |
| ACinf | adverb | conjunct | inferential | *otherwise* |
| AClis | adverb | conjunct | listing | *in addition* |
| ACres | adverb | conjunct | resultive | *consequently* |
| ACsum | adverb | conjunct | summative | *all in all* |
| ACtra | adverb | conjunct | transitional | *by the way* |
| ADcnt | adverb | disjunct | content | *probably* |
| ADsty | adverb | disjunct | style | *personally* |
| AEels | adverb | postmodifier | *else* | |
| ANnot | adverb | negative | *not* | |
| AQapo | adverb | discourse item | apology | *I'm sorry* |
| AQexp | adverb | discourse item | expletive | *fuck off* |
| AQgre | adverb | discourse item | greeting | *goodbye* |
| AQneg | adverb | discourse item | negative | *no* |
| AQord | adverb | discourse item | order | *give over* |
| AQpol | adverb | discourse item | politeness | *please* |
| AQpos | adverb | discourse item | positive | *yes* |
| AQres | adverb | discourse item | response | *I see* |
| AQsof | adverb | discourse item | softener | *I mean* |
| AQsup | adverb | discourse item | support | *[m]* |
| AQtha | adverb | discourse item | thanks | *thank you* |
| AQwel | adverb | discourse item | *well* | |

| | | | | |
|---|---|---|---|---|
| ASemp | adverb | subjunct | emphasizer | *actually* |
| ASfoc | adverb | subjunct | focusing | *mainly* |
| ASint | adverb | subjunct | intensifier | *a bit* |
| AXexi | adverb | existential | *there* | |
| AXexi*VB+3 | adverb | existential | *there's* | |
| AXexi*VM+8 | adverb | existential | *there'll* | |
| AXexi*VM+9 | adverb | existential | *there'd* | |
| B1ass | determiner | *of*-quantifier | assertive | *some of* |
| B1car | determiner | *of*-quantifier | cardinal | *one of* |
| B1mul | determiner | *of*-quantifier | multal | *more of* |
| B1neg | determiner | *of*-quantifier | negative | *little of* |
| B1non | determiner | *of*-quantifier | nonassertive | *any of* |
| B1ope | determiner | *of*-quantifier | open-class | *a couple of* |
| B1pau | determiner | *of*-quantifier | paucal | *few of* |
| B1qua | determiner | *of*-quantifier | quantitative | *a bit of* |
| B1uni | determiner | *of*-quantifier | universal | *all of* |
| B2deg | determiner | predeterminer | degree | *quite* |
| B2mul | determiner | predeterminer | multal | *once* |
| B2pau | determiner | predeterminer | paucal | *less* |
| B2qua | determiner | predeterminer | quantitative | *double* |
| B2uni | determiner | predeterminer | universal | *all* |
| B3ass | determiner | central | assertive | *some* |
| B3def | determiner | central | definite | *the* |
| B3deg | determiner | central | degree | *what* |
| B3dem | determiner | central | demonstrative | *that* |
| B3ind | determiner | central | indefinite | *a, an* |
| B3itr | determiner | central | interrogative | *whatever* |
| B3mul | determiner | central | multal | *much* |
| B3neg | determiner | central | negative | *neither* |
| B3non | determiner | central | nonassertive | *any* |
| B3pos | determiner | central | possessive | *your* |
| B3qua | determiner | central | quantitative | *enough* |
| B3uni | determiner | central | universal | *each* |
| B3rel | determiner | central | relative | *whose* |
| B4car | determiner | postdeterminer | cardinal | *five* |
| B4gor | determiner | postdeterminer | general ordinal | *additional* |
| B4mul | determiner | postdeterminer | multal | *many* |
| B4neg | determiner | postdeterminer | negative | *little* |
| B4oth | determiner | postdeterminer | *other*-type | *another* |

| | | | | |
|---|---|---|---|---|
| B4pau | determiner | postdeterminer | paucal | *few* |
| B4qua | determiner | postdeterminer | quantitative | *several* |
| BHass | pronoun | head | assertive | *some* |
| BHdem | pronoun | head | demonstrative | *that* |
| BHdem*VB+3 | pronoun | head | demonstrative | *that's* |
| BHdem*VM+8 | pronoun | head | demonstrative | *that'll* |
| BHgen | pronoun | head | general | *one* |
| BHitr | pronoun | head | interrogative | *what* |
| BHitr*VB+3 | pronoun | head | interrogative | *who's* |
| BHmul | pronoun | head | multal | *many* |
| BHneg | pronoun | head | negative | *no one* |
| BHneu | pronoun | head | personal | *it* |
| BHneu*VB+3 | pronoun | head | personal | *it's* |
| BHneu*VM+8 | pronoun | head | personal | *it'll* |
| BHneu*VM+9 | pronoun | head | personal | *it'd* |
| BHnon | pronoun | head | nonassertive | *any* |
| BHobj | pronoun | head | personal: obj | *me* |
| BHoth | pronoun | head | *other*-type | *another* |
| BHpau | pronoun | head | paucal | *few* |
| BHper | pronoun | head | personal | *you* |
| BHper*VB+4 | pronoun | head | personal | *you're* |
| BHper*VH+0 | pronoun | head | personal | *you've* |
| BHper*VM+8 | pronoun | head | personal | *you'll* |
| BHper*VM+9 | pronoun | head | personal | *you'd* |
| BHpos | pronoun | head | possessive | *hers* |
| BHqua | pronoun | head | quantitative | *enough* |
| BHrec | pronoun | head | reciprocal | *each other* |
| BHref | pronoun | head | reflexive | *herself* |
| BHrep | pronoun | head | replacive | *one* |
| BHsub | pronoun | head | personal: subj | *I* |
| BHsub*VB+1 | pronoun | head | personal+'*m* | *I'm* |
| BHsub*VB+3 | pronoun | head | personal+'*s* | *he's* |
| BHsub*VB+4 | pronoun | head | personal+'*re* | *they're* |
| BHsub*VH+0 | pronoun | head | personal+'*ve* | *I've* |
| BHsub*VM+8 | pronoun | head | personal+'*ll* | *I'll* |
| BHsub*VM+9 | pronoun | head | personal+'*d* | *I'd* |
| BHuni | pronoun | head | universal | *all* |
| CAcoo | conjunction | coordinator | *and* | |
| CBcoo | conjunction | coordinator | *but* | |

| | | | | |
|---|---|---|---|---|
| CCcom | conjunction | subordinator | comparative | *as, like* |
| CCcon | conjunction | subordinator | condition | *although* |
| CCexc | conjunction | subordinator | exception | *except* |
| CCitr | conjunction | subordinator | interrogative | *if, whether* |
| CCpre | conjunction | subordinator | preference | *rather than* |
| CCpro | conjunction | subordinator | proportion | *as* |
| CCpur | conjunction | subordinator | purpose | *lest* |
| CCrea | conjunction | subordinator | reason | *as* |
| CCres | conjunction | subordinator | result | *in order that* |
| CCspa | conjunction | subordinator | space, place | *as far as* |
| CCtim | conjunction | subordinator | time | *according as* |
| CDnom | conjunction | subordinator | *that* | |
| CRcoo | conjunction | coordinator | *or* | |
| FAcar | numeral | cardinal | | *eight, 8* |
| FBord | numeral | ordinal | | *eighth, 8th* |
| FCnom | numeral | nominal | unmarked | *hundred, 100* |
| FDnom | numeral | nominal | currency etc | *£1000* |
| GAwhi | pronoun | relative | *which* | |
| GAwho | pronoun | relative | *who* | |
| GAwho*VB+3 | pronoun | relative | *who+s* | |
| GAwho*VH+0 | pronoun | relative | *who+'ve* | |
| GAwho*VM+9 | pronoun | relative | *who+'d* | |
| GBwhm | pronoun | relative | *whom* | |
| GCwha | pronoun | relative | *what* | |
| GCwha*VB+3 | pronoun | relative | *what+'s* | |
| GDtha | pronoun | relative | *that* | |
| GDtha*VB+3 | pronoun | relative | *that+'s* | |
| GDtha*VM+8 | pronoun | relative | *that+'ll* | |
| GDtha*VM+9 | pronoun | relative | *that+'d* | |
| GEas | pronoun | relative | *as* | |
| JA | adjective | unmarked | *big* | |
| JA+R | adjective | comparative | | *bigger* |
| JA+T | adjective | superlative | | *biggest* |
| JH | adjective as NP head | | | (the) *rich* |
| JL | adjective | language | | (in) *Dutch* |
| JN | adjective | nationality | | *Dutch* (books) |
| NC | noun | common | unmarked | *church* |
| NC+2 | noun | common | plural | *academics* |
| NC+Z | noun | common | genitive | *hill's* |

| | | | | |
|---|---|---|---|---|
| NP | noun | proper | unmarked | (an) *Indian* |
| NP+2 | noun | proper | plural | (the) *Indians* |
| NP+Z | noun | proper | genitive | *Churchill's* |
| NX | nominal | abbreviation | | *BBC* |
| PA | preposition | | | *into* |
| TO | inf marker | | *to* | |
| VA+0 | verb | main | base | *add* |
| VA+0DN | verb | main | base/past/ppl | *broadcast* |
| VA+0N | verb | main | base/ppl form | *become* |
| VA+3 | verb | main | s-form | *becomes* |
| VA+D | verb | main | past form | *became* |
| VA+DN | verb | main | past/ppl form | *added* |
| VA+G | verb | main | ing-form | *adding* |
| VA+N | verb | main | ppl form | *taken* |
| VB+0 | verb | form of BE | *be* | |
| VB+1 | verb | form of BE | *am* | |
| VB+3 | verb | form of BE | *is* | |
| VB+3*ANnot | verb | form of BE | *isn't* | |
| VB+4 | verb | form of BE | *are* | |
| VB+4*ANnot | verb | form of BE | *aren't* | |
| VB+5 | verb | form of BE | *was* | |
| VB+5*ANnot | verb | form of BE | *wasn't* | |
| VB+6 | verb | form of BE | *were* | |
| VB+6*ANnot | verb | form of BE | *weren't* | |
| VB+G | verb | form of BE | *being* | |
| VB+N | verb | form of BE | *been* | |
| VD+0 | verb | form of DO | *do* | |
| VD+0*ANnot | verb | form of DO | *don't* | |
| VD+0*BHper | verb | form of DO | *d'you* | |
| VD+3 | verb | form of DO | *does* | |
| VD+3*ANnot | verb | form of DO | *doesn't* | |
| VD+D | verb | form of DO | *did* | |
| VD+D*ANnot | verb | form of DO | *didn't* | |
| VD+G | verb | form of DO | *doing* | |
| VD+N | verb | form of DO | *done* | |
| VH+0 | verb | form of HAVE | *have* | |
| VH+0*ANnot | verb | form of HAVE | *haven't* | |
| VH+3 | verb | form of HAVE | *has* | |
| VH+3*ANnot | verb | form of HAVE | *hasn't* | |

| | | | | |
|---|---|---|---|---|
| VH+DN | verb | form of HAVE | *had* | |
| VH+D*ANnot | verb | form of HAVE | *hadn't* | |
| VH+G | verb | form of modal | *having* | |
| VM+0*ANnot | verb | form of modal | *mustn't* | |
| VM+8*BHobj | verb | form of modal | *let's* | |
| VM+8 | verb | form of modal | *can* | |
| VM+8*ANnot | verb | form of modal | *can't* | |
| VM+9 | verb | form of modal | *could* | |
| VM+9*ANnot | verb | form of modal | *couldn't* | |
| VM+9*VH+0 | verb | form of modal | *should've* | |
| VMidi | verb | modal | idiom | *had better* |
| VMmar | verb | modal | marginal | *dare to* |
| VXcat | verb | catenative | | *appear to* |
| VXsem | verb | semi-auxiliary | | (be) *able to* |
| XX | misc | foreign words | | *en route* |

# 4

# An automatic word class tagger and a phrase parser

*Mats Eeg-Olofsson*

## 4.1 Introduction

The automatic word class tagging and parsing that will be described here have been used within the TESS project as preparatory steps for automatic segmentation.

The input to the word class tagger has consisted of machine-readable texts from the Brown Corpus. The resulting output is the same texts enriched with word class tags attached to the word tokens. The tags are needed in the subsequent syntactic parsing, where an automatic phrase parser analyses the tagged input as a sequence of syntactic phrases.

The output from the phrase parser is a labelled bracketing of the input, including the original words and word class tags. This bracketing serves as input to the automatic segmentation procedure, which implements grammatically based rules for segmentation of text into tone units.

## 4.2 An automatic word class tagger

The immediate purpose of constructing the word class tagger for TESS was to provide suitable input for the phrase parser, but the tagger could also serve as a tool for various experiments in automatic word class tagging.

While the general approach to word class tagging in some preceding research projects in Lund (see, for example, Svartvik et al 1982) is very similar to that of TESS, being essentially probabilistic and based on tag statistics, these projects differ in several ways from the specifications in TESS, in particular in terms of input, tagging system, and statistics:

(a) *Input*: previously the material was from LLC with origin in speech, now it was written material from the Brown Corpus;
(b) *Tagging system*: previously the word class tagging system was more coarse-grained, lacking many of the semantic distinctions of the new one (see Chapter 3, pp. 92ff);
(c) *Available statistics*: previously the statistics were based on some 50 000 tagged running words (10 LLC texts), now on virtually none.

The texts to be tagged for word class were editorials from various American newspapers (texts B01-B03 of the Brown Corpus, each comprising about 2000 running words). It was a great help in the initial stage of the tagging to have access to the tagged machine-readable version of the Brown Corpus. Systematic translation of the extant Brown tags in our little test corpus could be used as a first approximation to tagging according to the TESS tagset. This provided a practical solution to the bootstrap problem of statistical tagging: it presupposes the availability of statistics, which, in turn, have to be compiled from previously tagged texts.

The word class tagging system used in the experiments was an early version of the new tagset, the main difference being that the catenative and special modal verb categories (VMidi, VMmar, VXcat, VXsem) had not yet been introduced.

The statistical tagging algorithm used in the experiment was a variation of that described in my outline (Eeg-Olofsson 1985), which is very similar to that developed independently within the LOB tagging project (see, for example, Marshall 1983). With the present approach, word class tagging is regarded as the statistical problem of finding the most likely sequence of tags corresponding to a stretch of text that has been found to conform to certain graphic patterns. Such a pattern is characterized by the presence of features such as certain words, endings, initial capitals, etc, and associated with information about the relative frequencies with which various tags are assigned to the text matching the pattern. To compute the probability of a given tag sequence, the algorithm also uses a table of estimated transitional probabilities, containing information about, for instance, the probability that a word tagged as a noun is followed by another noun, etc.

Thus, the categories that can be kept apart by the use of this algorithm are mainly such as are associated with surface graphic patterns and/or short-range syntactic properties. In view of this fact, the TESS tagset would seem to be both easier and yet more difficult to computerize than previous systems. The disappearance of the distinction between past (tag VA+D) and past participle (tag VA+N) for regular *-ed* forms (now tagged VA+DN) has undoubtedly made the task easier. On the other hand, the semantically based subclassification of adverbs (AApro, AAspa, ACapp, ADsty, etc) and conjunctions (CAcoo, CCexc, CCpur, etc) has introduced new difficulties for the algorithm.

The statistical background data used in the experiments consisted of a dictionary, a list of suffixes, and a tag transition table. The dictionary, containing some 2200 entries, was based on ten texts from LLC, totalling about 50 000 word tokens, supplemented with additional entries for closed-class categories, such as prepositions, irregular verbs, etc, as well as for the adverbs in the test corpus. The suffix list contained some 500 entries, based on those published by Johansson & Jahr (1982). The probabilities in the tag transition table were estimated from a manually tagged version of the test corpus.

The error rate, computed by comparing the tags predicted by the computer algorithm with those assigned manually, varied between 3 and 6 per cent of the running words. Hence, it compares well with the results reported by Leech, Garside & Atwell (1983). Experiments in varying the amount of statistical information associated with the entries in the dictionary and the suffix list produced error rates differing only by one or two percentage points.

One error type was the confusion between word classes that can have the same endings, for example *-ing* in nouns (*The singing was loud*), adjectives (*a frightening experience*), and verbs (*You are frightening the children*) and *-al* in nouns (*He's a criminal*) and adjectives (*criminal behaviour*). Another common type of error occurred when third person singular present verb forms (as in *she walks barefoot*) were mistaken for plural nouns with the same *-s* ending (as in *long country walks*), probably because the latter category is more frequent.

Among the words contained in the dictionary, a few such as *some* and *that* were frequently mistagged as determiners when functioning as pronouns. Compare:

| | |
|---|---|
| *Some* people say that ... | *Some* say that ... |
| *That* was my mistake. | *That* mistake was mine. |

A similar type of error afflicted words like *before* identified as prepositions instead of conjunctions. Compare:

| | |
|---|---|
| She left before John. | [preposition] |
| She left before John came. | [conjunction] |

The individual words that caused the most trouble were the following which are all both frequent and multi-functional: *as* (see pp 87f), *that* (determiner, pronoun, conjunction, or intensifier), and *to* (which may be either preposition or infinitive marker in ... *ways to finance* ...).

The semantic subclassification of adverbs and conjunctions caused fewer errors than expected. Whether this will be true for larger corpora as well remains to be seen. The good result in this case may be a consequence of the fact that the tag statistics used were based on the test corpus itself.

In summary, the results, while far from being conclusive, suggest that the statistical approach is useful for tagging text according to the TESS tagset. It is evident that further refinement of this basic approach will produce only marginal improvement of the success rate. Methods using a small text 'window' will always have great difficulty in distinguishing between, eg prepositions and conjunctions, where the decision must be based on the examination of a larger syntactic context. Since problems such as these must be left to human post-editors, future work on designing efficient systems for tagging large corpora should concentrate not so much on optimizing computer algorithms as on the man-machine interaction within the system. Furthermore, the accumulated knowledge about what statistical word class tagging can and cannot do should be exploited when plans are made to analyse text corpora both syntactically (at constituent level) and at word class level (mainly as a preparation for syntactic analysis).

## 4.3 An automatic phrase parser

### 4.3.1 Introduction

This is a description of the experimental phrase parser used in the TESS project. It is written in the Definite Clause Grammar formalism implemented in the Prolog programming language. [Note: In the listing below, program code is indented; comments in the code are preceded by the percentage sign %; running text starts in the left-hand margin.]

The parser is implemented in Sussex Prolog, a Prolog dialect which is included in the POPLOG programming package and has a syntax that is very close to the inofficial Edinburgh standard. The parsing program has been run on a VAX-11/730 computer under the VAX/VMS operating system.

**Figure 4:1.** Input example from text B01 in the Brown Corpus.

| WORD | TESS TAG | BROWN TAG | LOCATION REFERENCE |
|---|---|---|---|
| ASSEMBLY | NC | NN-HL | B01001001E1 |
| SESSION | NC | NN-HL | B01001002E1 |
| BROUGHT | VA+DN | VBD-HL | B01001003E1 |
| MUCH | B3mul | AP-HL | B01001004E1 |
| GOOD | NC | NN-HL | B01001005E1 |
| THE | B3def | AT | B01001006E1 |
| *GENERAL | NP/22 | NP | B01001007E1 |
| *ASSEMBLY | NP/21 | NP | B01001008E1 |
| , | , | , | B01001009E1 |
| WHICH | GAwhi | WDT | B01001010E1 |
| ADJOURNS | VA+3 | VBZ | B01002001E1 |
| TODAY | AAtim | NR | B01002002E1 |
| , | , | , | B01002003E1 |
| HAS | VH+3 | HVZ | B01002004E1 |
| PERFORMED | VA+DN | VBN | B01002005E1 |
| IN | PA | IN | B01002006E1 |
| AN | B3ind | AT | B01002007E1 |
| ATMOSPHERE | NC | NN | B01002008E1 |
| OF | PA | IN | B01002009E1 |
| CRISIS | NC | NN | B01002010E1 |
| AND | CAcoo | CC | B01002011E1 |
| STRUGGLE | NC | NN | B01002012E1 |
| FROM | PA | IN | B01003001E1 |
| THE | B3def | AT | B01003002E1 |
| DAY | NC | NN | B01003003E1 |
| IT | BHneu | PPS | B01003004E1 |
| CONVENED | VA+DN | VBD | B01003005E1 |
| . | . | . | B01003006E1 |

The input to the parser consists of a tagged Brown Corpus text in 'vertical' format, with one word per line, including word class tags (both the original Brown tag and the TESS tag) and a location reference. The Brown tags are useful for separating headlines from the rest of the text.

The input example in Figure 4:1 consists of the first headline and the first following period of text B01.

The input routine of the parser divides the text into period segments, each consisting of either a sequence of words terminated by a major punctuation mark, or a headline. Then the parser transforms the input segment into a list of tagged grammatical words, splitting up contractions into their components (eg *DON'T* into *do* and *not*), assembling ditto-tagged words into units (eg **GENERAL* and **ASSEMBLY* into *General Assembly*), and changing capitalization in accordance with normal orthographic practice.

**Figure 4:2.** The 'display' version of the output of the syntactic analysis.

```
===============================================================
B01001001E1 -- B01001005E1
assembly session brought much good
===============================================================
NPH: [nmod, nhead]
   nmod:
      ASSEMBLY                          NC            B01001001E1
   head:
      SESSION                           NC            B01001002E1
VPH:
   head:
      BROUGHT                           VA+DN         B01001003E1
NPH: [det, nhead]
   det:
      MUCH                              B3mul         B01001004E1
   head:
      GOOD                              NC            B01001005E1

================================================================
B01001006E1 -- B01003006E1
the General Assembly , which adjourns today , has performed in
an atmosphere of crisis and struggle from the day it convened .
================================================================
NPH: [det, nhead]
   det:
      THE                               B3def         B01001006E1
   head:
      *GENERAL *ASSEMBLY                NP/2          B01001007E1
([,, ,, ,, B01001009E1])
NPH: [rel]
   head:
      WHICH                             GAwhi         B01001010E1
VPH:
   head:
      ADJOURNS                          VA+3          B01002001E1
APH:
   head:
      TODAY                             AAtim         B01002002E1
([,, ,, ,, B01002003E1])
VPH: [perf]
   op:
      HAS                               VH+3          B01002004E1
   head:
      PERFORMED                         VA+DN         B01002005E1
PPH:
   prep:
      IN                                PA            B01002006E1
   head:
      nph: [det, nhead, postmod]
         head: [det, nhead]
            det:
               AN                       B3ind         B01002007E1
            head:
               ATMOSPHERE               NC            B01002008E1
         postmod:
            prep:
```

**Figure 4:2**, contd

```
                 OF                     PA             B01002009E1
              head:
                 nph: [nhead]
                    head:
                       CRISIS           NC             B01002010E1
([AND, CAcoo, CC, B01002011E1])
NPH: [nhead]
   head:
      STRUGGLE                          NC             B01002012E1
PPH:
   prep:
      FROM                              PA             B01003001E1
   head:
      nph: [det, nhead]
         det:
            THE                         B3def          B01003002E1
         head:
            DAY                         NC             B01003003E1
NPH: [pro]
   head:
      IT                                BHneu          B01003004E1
VPH:
   head:
      CONVENED                          VA+DN          B01003005E1
([., ., ., B01003006E1])
```

This list of tagged grammatical words is the input to the analysis proper, which parses it as a sequence of phrases of various types. The output of the analysis is a labelled bracketing, which can be displayed in different forms for visual inspection or further processing by computer.

Figure 4:2 shows the 'display' version of the output of the syntactic analysis of the two file segments from text B01 in Figure 4:1.

The output version of the same two segments that will be used as input by the segmentation program is shown in Figure 4:3.

The parser does a good job of splitting up the text into grammatical phrases, considering the very small amount of lexical and higher-level syntactic information that is available to it. Its main weaknesses lie in the handling of scope ambiguities, particularly in coordinated constructions. The segmentation rules (see Chapter 12) have been designed so as to compensate for these shortcomings. The parser is still very much an experimental one, whose performance can be improved in innumerable minor respects. The following description should be seen as a sketch of some basic ideas, rather than a definitive solution.

**Figure 4:3**. The output version used as input by the segmentation program.

```
[
  nph(
    [
      nmod(
        [
          word(['ASSEMBLY', 'NC', 'NN-HL', 'B01001001E1'])
        ],
        []
      ),
      head(
        [
          word(['SESSION', 'NC', 'NN-HL', 'B01001002E1'])
        ],
        []
      )
    ],
    [
      nmod,
      nhead
    ]
  ),
  vph(
    [
      head(
        [
          word(['BROUGHT', 'VA+DN', 'VBD-HL', 'B01001003E1'])
        ],
        []
      )
    ],
    []
  ),
  nph(
    [
      det(
        [

          word(['MUCH', 'B3mul', 'AP-HL', 'B01001004E1'])
        ],
        []
      ),
      head(
        [
          word(['GOOD', 'NC', 'NN-HL', 'B01001005E1'])
        ],
        []
      )
    ],
    [
      det,
      nhead
    ]
  )
].
[
  nph(
    [
      det(
        [
```

**Figure 4:3**, contd

```
            word(['THE', 'B3def', 'AT', 'B01001006E1'])
         ],
         []
      ),
      head(
         [
            word(
               [
                  '*GENERAL *ASSEMBLY',
                  'NP/2',
                  'NP',
                  'B01001007E1',
                  'NP',
                  'B01001008E1'
               ]
            )

         ],
         []
      )
   ],
   [
      det,
      nhead
   ]
),
other([',', ',', ',', 'B01001009E1']),
nph(
   [
      head(
         [
            word(['WHICH', 'GAwhi', 'WDT', 'B01001010E1'])
         ],
         []
      )
   ],
   [
      rel
   ]
),
vph(
   [
      head(
         [
            word(['ADJOURNS', 'VA+3', 'VBZ', 'B01002001E1'])
         ],
         []
      )
   ],
   []
),
aph(
   [
      head(
         [

            word(['TODAY', 'AAtim', 'NR', 'B01002002E1'])
         ],
         []
      )
   ],
   []
```

**Figure 4:3**, contd

```
),
other([',', ',', ',', 'B01002003E1']),
vph(
   [
      op(
         [
            word(['HAS', 'VH+3', 'HVZ', 'B01002004E1'])
         ],
         []
      ),
      hole(
         nil
      ),
      head(
         [
            word(['PERFORMED', 'VA+DN', 'VBN', 'B01002005E1'])
         ],
         []
      )
   ],
   [
      perf
   ]
),
pph(
   [
      prep(
         [
            word(['IN', 'PA', 'IN', 'B01002006E1'])
         ],
         []
      ),

      head(
         [
            nph(
               [
                  head(
                     [
                        det(
                           [
                              word(['AN', 'B3ind', 'AT', 'B01002007E1'])
                           ],
                           []
                        ),
                        head(
                           [
                              word(['ATMOSPHERE', 'NC', 'NN', 'B01002008E1'])
                           ],
                           []
                        )
                     ],
                     [
                        det,
                        nhead
                     ]
                  ),
                  postmod(
                     [
                        prep(
                           [
                              word(['OF', 'PA', 'IN', 'B01002009E1'])
```

**Figure 4:3**, contd

```
                        ],
                        []
                      ),
                      head(
                        [
                          nph(
                            [
                              head(
                                [
                                  word(['CRISIS', 'NC', 'NN',
                                  'B01002010E1'])
                                ],
                                []
                              )
                            ],
                            [
                              nhead
                            ]
                          )
                        ],
                        []
                      )
                    ],
                    []
                  )
                ],
                [
                  det,
                  nhead,
                  postmod
                ]
              )
            ],
            []
          )
        ],
        []
      ),
      other(['AND', 'CAcoo', 'CC', 'B01002011E1']),
      nph(
        [
          head(
            [
              word(['STRUGGLE', 'NC', 'NN', 'B01002012E1'])
            ],
            []
          )
        ],

        [
          nhead
        ]
      ),
      pph(
        [
          prep(
            [
              word(['FROM', 'PA', 'IN', 'B01003001E1'])
            ],
            []
          ),
```

**Figure 4:3**, contd

```
        head(
          [
            nph(
              [
                det(
                  [
                    word(['THE', 'B3def', 'AT', 'B01003002E1'])
                  ],
                  []
                ),
                head(
                  [
                    word(['DAY', 'NC', 'NN', 'B01003003E1'])
                  ],
                  []
                )
              ],
              [
                det,
                nhead
              ]
            )
          ],
          []
        )
      ],

      []
    ),
    nph(
      [
        head(
          [
            word(['IT', 'BHneu', 'PPS', 'B01003004E1'])
          ],
          []
        )
      ],
      [
        pro
      ]
    ),
    vph(
      [
        head(
          [
            word(['CONVENED', 'VA+DN', 'VBD', 'B01003005E1'])
          ],
          []
        )
      ],
      []
    ),
    other(['.', '.', '.', 'B01003006E1'])
  ].
```

### 4.3.2 Parsing strategy and top-level predicates

The top-level parsing predicate ANALYSIS instantiates its only argument to a structural description of the text segment ('sentence') to be parsed. The description will consist of a list of structural descriptions of phrases. ANALYSIS uses the parsing predicate PHRASE to parse phrases (constituents) of various kinds. The parsing is non-deterministic and can produce different solutions on backtracking, but only the first parse is output for further processing (segmentation). Some care has been taken to make this selected parse as good as possible. The basic strategy is to try and make the component phrases as inclusive as possible, so that they contain a maximum number of grammatically compatible words. A practical disadvantage of this method is that it necessitates a large amount of backtracking, which makes the parser just a little too slow for interactive use.

```
analysis([Phrase1|Phrases]) -->
    phrase(Phrase1), % Parse a phrase
    pairanalysis(Phrase1,Phrases).
    % Go on, checking compatibility
```

Basically, ANALYSIS parses a sentence as an unstructured sequence of phrases. The auxiliary predicate PAIRANALYSIS, which checks the compatibility of successive phrases, makes it possible to add some structure to this sequence. PAIRANALYSIS uses the basic parsing predicate PHRASE and the auxiliary predicate COMPATIBLE_PHRASES. Its main use here is as a kind of post-mortem device to improve the analysis of certain phrases containing *-ing* and *-ed* forms. This is a consequence of the fact (referred to in Section 12.3) that such forms are tagged on the word-class level for form rather than for function. PAIRANALYSIS, or some extension, might also be used to improve the handling of coordinated constructions, which are not analysed as such by the present parser.

```
pairanalysis(Anyphrase,[],[],[]). % Nothing left
pairanalysis(Phrase1,[Phrase2|Phrases]) -->
    phrase(Phrase2), % Parse a new phrase
    { compatible_phrases(Phrase1,Phrase2) },
    % Check compatibility
    pairanalysis(Phrase2,Phrases).
    % Go on recursively
compatible_phrases(P1,P2) :-
    P1 = nph(_,F1), infeatlist(inghead,F1), !,
    anyfeat([gen,postdet,det],F1),
    ( infeatlist(postmod,F1) ;
    P2 = pph(_,_) ).
```

If an *-ing* form is the head of a noun phrase (feature INGHEAD as in *the shooting of the hunters*), it must be preceded by a determiner, postdeterminer or genitive, and followed by a preposition either in the same noun phrase (POSTMOD) or in the following phrase.

```
compatible_phrases(_,_).
    % Catchall -- all other phrase pairs are compatible
```

### 4.3.3 Types of phrases

PHRASE is the basic parsing predicate, describing various types of constituents: verb phrases (of two kinds), prepositional phrases, adjective phrases, noun phrases, adverb phrases, and 'others' that cannot be analysed in these terms, eg punctuation marks and conjunctions.

```
phrase(Desc) -->
    vph_noning(Desc).
    % Verb phrase not introduced by -ing form
phrase(Desc) -->
    pph(Desc,_). % Prepositional phrase
phrase(nph(D,F)) -->
    nph(nph(D,F)),
    % Noun phrase with feature list F
    {( nofeats([ingmod],F) ;
    infeatlist(ingmod,F),
    anyfeat([gen,det,postdet],F) ),
    nofeats([edmod],F)}.
```

Filter: *-ing* forms are accepted as premodifiers only if preceded by a determiner, postdeterminer, or genitival noun phrase; *-ed* forms are accepted as premodifiers only in prepositional phrases.

```
phrase(Desc) -->
    jph(Desc). % Adjective phrase
phrase(Desc) -->
    aph(Desc,_,_). % Adverb phrase
phrase(Desc) -->
    vph_ing(Desc).
    % Verb phrase introduced by -ing form
phrase(Desc) -->
    other(Desc). % 'Ragbag'
```

The various phrase predicates have the overall format:

```
Name(Name(Daughters,Features), Feature1, Feature2 ... )
```

where Name is any of the phrase types aph, jph, nph, vph, and pph. (The first occurrence of Name is the name of the predicate, while the second is a label for the structural description.) Daughters is a role list, constituting a structural description of the phrase. Features is a (usually empty) list of descriptive features. Feature1, Feature2, ... are arguments used by the parsing routines only and are not parts of the structural description. Entries of the role lists have one of the following formats:

```
Name(Daughters,Features)
Function(Daughters,Features)
word(Word)
other(Word)
hole(Name(Daughters,Features))
hole(nil).
```

The first two formats are used for subconstituents. As above, Name is the name of some phrase type. Function is any of the following words:

```
int,head,postmod,gen,quant,predet,det,postdet,
adjmod,nmod,prep,op,tail,compl,cop,part,infmark,coord
```

which describe the function of the subconstituent in the superordinate constituent.

The following WORD format describes grammatical terminals, where Word is a list of the form [Body,Tag,Browntag,Location], corresponding to an input word in the vertical format illustrated above.

The OTHER format describes words that cannot be analysed in terms of the above phrase types.

The HOLE format is used to describe discontinuous phrases, like the verb phrase in *did he look it up*, which is analysed as a discontinuous verb phrase, consisting of an operator, a main verb and a particle, in addition to two 'holes', filled by the noun phrases *he* and *it*, respectively. The format hole(nil) can be used to describe unfilled potential holes in phrases.

The structural description of the top level parsing predicate ANALYSIS is a role list, where all entries are either of the type 'Name' or of the type 'other'.

### 4.3.4 Description of adverb phrase

The predicate APH parses adverb phrases. It has three arguments: structural description, modification (simple/enough/intens), and head type (intens/nonintens).

```
aph(aph([int([word(X)|Y],[]), Z], []),
    intens,Type) -->
    [X], {intensifier(X)},
    aph(aph([int(Y,[]), Z], []), intens, Type).
    % Several intensifiers, eg much more easily

aph(aph([int([word(X)],[]), Z], []),
    intens, Type) -->
```

```
    [X], {intensifier(X)},
    aph(aph([Z], []), simple, Type).
    % Nearest intensifier

aph(aph([Head,int([word(E)],[])], []), enough,
    Type) -->
    aph(aph([Head],[]), simple, Type),
    [E], {word(E,enough), adverb(E)}.
    % Adverb + enough, eg quickly enough

aph(aph([head([word(X)],[])], []), simple,
    intens) -->
    [X], {intensifier(X)}.
    % Intensifier as head, eg completely

aph(aph([head([word(X)],[])], []), simple,
    nonintens) -->
    [X], {adverb(X), not(intensifier(X))}.
    % Non-intensifier as head, eg clearly
```

Description of adverbs, including intensifiers:

```
adverb(X) :-
    has_tag(X,Tag), prefix(Tag,'A').
    % Word class tag begins with A

intensifier(X) :-
    has_tag(X,Tag), prefix(Tag,'ASint').
    % Word class tag ASint
```

### 4.3.5 Description of adjective phrase

The predicate JPH, which parses adjective phrases, has one argument: a structural description of the form jph(Struct,[]). Several consecutive adjectives are analysed as a sequence of adjective phrases.

```
jph(jph([int([X],[]), head([word(Y)],[])], [])) -->
    % Intensified, eg very good
    aph(X,_,intens),
    [Y], { adjectival(Y) }.

jph(jph([head([word(X)],[]), int([word(Y)],[])],[])) -->
    % With enough, eg good enough.
    [X], { adjectival(X) },
    [Y], { word(Y,enough), intensifier(Y) }.

jph(jph([head([word(X)],[])], [])) -->
    % Single head -- adjective proper, eg good
    [X], {adjective(X)}.
```

The following predicates describe adjectives in terms of word class tags:

```
adjective(X) :-
    has_tag(X,Tag), prefix(Tag,'J').
    % JA, JH, JN tags
```

```
adjective(X) :-
    has_tag(X,Tag), prefix(Tag,'FA'). % Cardinal

adjective(X) :-
    has_tag(X,Tag), prefix(Tag,'FB'). % Ordinal
```

The predicate ADJECTIVAL includes certain verb forms in addition to words tagged as adjectives:

```
adjectival(X) :-
    (adjective(X) ;
    % Adjective proper, as defined above
    psp(X) ;
    % Present participle -- see definition
    % in verb section
    ptp(X)). % Past participle
```

### 4.3.6 Description of noun phrases

The predicate NPH parses noun phrases. It has one argument of the form nph(Struct,Feats). NPH uses the predicate NPH_LEFT to parse the main part of the noun phrase. In addition, there can be a postmodifier or a preceding focussing adverb.

```
nph(nph([foc([W],[]),head([X],[])],[focus])) -->
    % With focussing adverb, eg even children
    aph(W,_,_), % Preceding adverb phrase
    { get_head_word(W,H), has_tag(H,T),
    prefix(T,'ASfoc') }, % Focussing head
    nph(X).
nph(nph([head(X,Y),postmod(Z,W)], V)) -->
    % With postmodifier, eg an atmosphere of crisis
    nph_left(nph(X,Y)), % NPH proper
    pph(pph(Z,W),Prep),
    % Prepositional phrase as postmodifier
    { ( Prep = of ; Prep = per ; Prep = a ) },
    { addfeatz(Y,postmod,V) }.
```

Prepositional phrases are attached as postmodifiers to noun phrases only if the preposition is *of*, *per* or *a* (*twice a day*). This is a heuristic partial solution to the classical problem of prepositional phrase attachment. Note, however, that it is not completely safe to attach a PPH introduced by *of* to an immediately preceding NPH, as seen in the example *acceptance by the state of responsibility for ...*

```
nph(X) --> nph_left(X).
    % No extensions to main part of noun phrase
```

The predicate NPH_LEFT uses the auxiliary predicate NPH_GEN to handle noun phrases containing genitive premodifiers. Other noun phrases are parsed by the predicate NPH_QUANT.

```
nph_left(X) --> % With genitive premodifier(s),
    % eg Georgia's financial problems
    { nph_gen_nesting(Y) },
    % Maximum nesting degree of genitives
    nph_gen(X,Y,_,_).

nph_left(X) --> nph_quant(X,_,_ ).
    % No genitive modifier, eg financial problems
```

In order to handle genitive premodifiers, which introduce left recursion, a threshold for the nesting of genitives is set at three:

```
nph_max_gen_nesting(3).
```

On backtracking, the predicate NPH_GEN_NESTING successively instantiates its argument to the permissible degrees of genitive nesting, starting with the maximum degree:

```
nph_gen_nesting(X) :-
    nph_max_gen_nesting(N),
    gennat(N,L),
    % Generate list of numbers N,...,1
    member(X,L). % Pick a number from list
```

The predicate NPH_GEN parses noun phrases containing genitive premodifiers. It has four arguments: a structural description of the form nph(Struct,Feats), the degree of nesting (1-3), the type of the head, and the case of the head noun.

```
nph_gen(nph([gen(X,Y),head(Z,W)], U), 1, Type, Case) -->
    % One genitive, eg Peter's girlfriend
    nph_quant(nph(X,Y),_,gen),
    nph_postdet(nph(Z,W),Type,Case),
    { addfeata(W,gen,U) }.

nph_gen(nph([gen(X,Y),head(Z,W)], U),Gens,Type,Case) -->
    % Nesting, eg Peter's girlfriend's car
    { Gens > 1, Gen1 is Gens - 1 },
    nph_gen(nph(X,Y),Gen1,_,gen),
    nph_postdet(nph(Z,W),Type,Case),
    { addfeata(W,gen,U) }.
```

The predicate NPH_QUANT parses quantifiers (if any), before calling NPH_PREDET. It has three arguments: a structural description of the form nph(Struct,Feats), the type of the head, and the case of the head. Only NPHs with nominal heads can have quantifiers.

```
nph_quant(nph([quant([word(Q)],[])|Premods],
    Nfeats), nhead, Case) -->
    [Q], { quant(Q) },
    nph_predet(nph(Premods, Ofeats), nhead, Case),
    { addfeata(Ofeats,quant,Nfeats) }.
    % Quantifier, eg plenty of that stuff
```

```
nph_quant(Struct, Type, Case) -->
    nph_predet(Struct, Type, Case).
    % No quantifiers
```

NPH_PREDET is like NPH_QUANT, but handles predeterminers instead of quantifiers, before calling NPH_DET.

```
nph_predet(nph([predet([word(NPd)|Pds],[])|Premods],
    Feats), nhead, Case) -->
    [NPd], { predet(NPd) },
    nph_predet(nph([predet(Pds,[])|Premods],
    Feats), nhead, Case).
    % Several predeterminers

nph_predet(nph([predet([word(Pd)],[])|Premods],
    Nfeats), nhead, Case) -->
    [Pd], { predet(Pd) },
    nph_det(nph(Premods, Ofeats), nhead, Case),
    { addfeata(Ofeats,predet,Nfeats) }.
    % Last predeterminer, eg such a nice dream

nph_predet(Struct, Type, Case) -->
    nph_det(Struct, Type, Case).
    % No predeterminer
```

Similarly, NPH_DET handles determiners, before calling NPH_POSTDET.

```
nph_det(nph([det([word(D)],[])|Premods], Nfeats),
    nhead, Case) -->
    [D], { det(D) },
    nph_postdet(nph(Premods, Ofeats), nhead, Case),
    { addfeata(Ofeats,det,Nfeats) }.
    % Determiner, eg its companion issue

nph_det(Struct, Type, Case) -->
    nph_postdet(Struct, Type, Case).
    % No determiner
```

NPH_POSTDET parses postdeterminers in the same way as the preceding predicates, before calling NPH_ADJMOD.

```
nph_postdet(nph([postdet([word(NPd)|Pds],[])|Premods],
    Feats), nhead, Case) -->
    [NPd], { postdet(NPd) },
    nph_postdet(nph([postdet(Pds,[])|Premods],
    Feats), nhead, Case).
    % Several postdeterminers

nph_postdet(nph([postdet([word(Pd)],[])|Premods],
    Nfeats), nhead, Case) -->
    [Pd], { postdet(Pd) },
    nph_adjmod(nph(Premods, Ofeats), nhead, Case),
    { addfeata(Ofeats,postdet,Nfeats) }.
    % Last postdeterminer, eg some other good books
```

```
nph_postdet(Struct, Type, Case) -->
    nph_adjmod(Struct, Type, Case).
    % No postdeterminer
```

NPH_ADJMOD attaches as many adjective phrases or 'adjectival' expressions as possible to the noun phrase before calling NPH_NMOD for noun-noun modifiers.

```
nph_adjmod(nph([adjmod([J|Js],F)|Premods], Feats),
    nhead, Case) -->
    jph(J),
    nph_adjmod(nph([adjmod(Js,F)|Premods], Feats),
    nhead, Case).
    % Adjective phrase, eg very good fruit

nph_adjmod(nph([adjmod([word(J)|Js],Fn)|Premods],
    Nfeats), nhead, Case) -->
    [J], { psp(J) },
    nph_adjmod(nph([adjmod(Js,F)|Premods], Feats),
    nhead, Case),
    { addfeata(F,ingmod,Fn),
    addfeata(Feats,ingmod,Nfeats) }.
    % -ing forms: additional checking in predicate
    % PHRASE, eg penetrating attention

nph_adjmod(nph([adjmod([word(J)|Js],Fn)|Premods],
    Nfeats), nhead, Case) -->
    [J], { ptp(J) },
    nph_adjmod(nph([adjmod(Js,F)|Premods], Feats),
    nhead, Case),
    { addfeata(F,edmod,Fn),
    addfeata(Feats,edmod,Nfeats) }.
    % Past participle as modifier: additional checking
    % in predicate PHRASE, eg balanced defense

nph_adjmod(nph([adjmod(Js,F)|Premods], Feats),
    nhead, Case) -->
    [Comma], { has_tag(Comma,',') },
    % Skip commas in lists of adjectives
    nph_adjmod(nph([adjmod(Js,F)|Premods], Feats),
    nhead, Case).

nph_adjmod(nph([adjmod([J],[])|Premods], Nfeats),
    nhead, Case) -->
    jph(J),
    nph_nmod(nph(Premods, Ofeats), nhead, Case),
    { addfeata(Ofeats,adjmod,Nfeats) }.
    % Nearest adjective phrase

nph_adjmod(nph([adjmod([word(J)],[ingmod])|Premods],
    Nfeats), nhead, Case) -->
    [J], { psp(J) },
    nph_nmod(nph(Premods, Ofeats), nhead, Case),
    { addfeata(Ofeats,ingmod,Nfeats) }.
    % Nearest -ing form
```

```
nph_adjmod(nph([adjmod([word(J)],[edmod])|Premods],
    Nfeats), nhead, Case) -->
    [J], { ptp(J) },
    nph_nmod(nph(Premods, Ofeats), nhead, Case),
    { addfeata(Ofeats,edmod,Nfeats) }.
    % Nearest PTP as premodifier

nph_adjmod(Struct, Type, Case) -->
    nph_nmod(Struct, Type, Case).
    % No more adjectives, PTP or -ing forms
```

NPH_NMOD takes care of noun-noun modifiers, before the head of the noun phrase is parsed by NPH_HEAD. Sequences of nouns are hard to parse correctly with this parser. In general, they will all be included in one single NPH. Clause-level information (eg about verb valency) is needed to split up such NPHs in the relevant cases.

```
nph_nmod(nph([nmod([word(N)|Ns],[])|Premods],
    Feats), nhead, Case) -->
    [N], { noun(N) },
    nph_nmod(nph([nmod(Ns,[])|Premods], Feats),
    nhead, Case).
    % Several noun modifiers,
    % eg road maintenance bond issue

nph_nmod(nph([nmod([word(N)],[])|Premods],
    Nfeats), nhead, Case) -->
    [N], { noun(N) },
    nph_head(nph(Premods, Ofeats), nhead, Case),
    { addfeata(Ofeats,nmod,Nfeats) }.
    % Last noun modifier, eg budget fight

nph_nmod(Struct, Type, Case) -->
    nph_head(Struct, Type, Case).
    % No more noun modifiers
```

NPH_HEAD parses the head of the noun phrase:

```
nph_head(nph([head([word(X)],[])],[pro]), pro, Case) -->
    [X], { headpron(X), case(X,Case) }.
    % NPH with pronoun as head, eg I, her

nph_head(nph([head([word(X)],[])],[rel]), rel, Case) -->
    [X], { relpron(X), case(X,Case) }.
    % Relative pronoun as head, eg who, which

nph_head(nph([head([word(X)],[])], [nhead]), nhead,
    Case) -->
    [X], { nhead(X), case(X,Case) }.
    % Nominal head, eg State, Rusk, two

nph_head(nph([head([word(X)],[])], [inghead]), nhead,
    Case) -->
    [X], { inghead(X), case(X,Case) }.
    % -ing form as head, eg racing
```

An *-ing* form can be the head of a noun phrase. Further checking is carried out in the predicate COMPATIBLE_PHRASES.

Descriptions of NPH head words in terms of word class tags:

```
headpron(X) :-
    % Pronouns that can be NPH heads
    has_tag(X,Tag),
    prefix(Tag,'BH'), % Tag must begin with BH
    not(prefix(Tag,'BHrep')). % But not BHrep

relpron(X) :- has_tag(X,Tag), prefix(Tag,'G').
    % Relative pronouns

nhead(X) :- has_tag(X,Tag), % Nominal heads
    ( prefix(Tag,'N') ; % Nouns
    prefix(Tag,'BHrep') ; % Replacive one
    prefix(Tag,'FDnom') ; % Nominal numerals
    prefix(Tag,'FAcar') ). % Cardinals

inghead(X) :- psp(X).
    % -ing forms can be NPH heads

noun(X) :- has_tag(X,Tag),
    % Nouns proper, including nominal numerals
    ( prefix(Tag,'N') ; prefix(Tag,'FDnom') ).
```

Descriptions of various kinds of determiners in terms of word class tags:

```
% Postdeterminers:
postdet(Z) :-
    has_tag(Z,Tag), prefix(Tag, 'B4').
% Tag begins with B4

% Determiners:
det(Z) :-
    has_tag(Z,Tag), prefix(Tag, 'B3').
% Tag begins with B3

% Predeterminers:
predet(Z) :-
    has_tag(Z,Tag), prefix(Tag, 'B2').
% Tag begins with B2

% Quantifiers:
quant(Z) :-
    has_tag(Z,Tag), prefix(Tag, 'B1').
% Tag begins with B1
```

Description of case in terms of word class tags:

```
case(X,gen) :- has_tag(X,'GEwhs'), !. % 'whose'

case(X,gen) :-
    has_tag(X,Tag), puretag(Tag,Ptag),
    suffix(Ptag,'+Z'). % Genitive tag
```

```
case(X,nom) :-
    has_tag(X,Tag), puretag(Tag,Ptag),
    not(suffix(Ptag,'+Z')).
```

### 4.3.7 Description of prepositional phrase

The predicate PPH parses prepositional phrases. It has two arguments: a structural description of the form pph(Struct,Feats) and the preposition.

```
pph(pph([prep([word(P)],[]),head([H],[])],[]),Pword) -->
    [P],
    { has_tag(P,Tag), prefix(Tag,'PA'),
    word(P,Pword) },
    nph(H). % eg in the schools
```

### 4.3.8 Verb phrase rules

The predicate VPH_NONING parses verb phrases not introduced by *-ing* forms. It instantiates its only argument to a structural description of the verb phrase of the form vph(Daughters,Features). This description is 'inherited' from the predicates VPH_DO, VPH_MODAL, VPH_INF, and VPH_PERF, which are called by VPH_NONING.

```
vph_noning(X) --> vph_do(X).
vph_noning(X) --> vph_modal(X,_).
vph_noning(X) --> vph_inf(X).
vph_noning(X) -->
    { finite(Y) }, vph_perf(X,Y).
    % Finite first verb
```

The predicate VPH_ING parses verb phrases introduced by *-ing* forms, by calling VPH_PERF. Such verb phrases are considered only after the parser has tried to match the input with all other types of phrases.

```
vph_ing(vph(X,Y)) -->
    vph_perf(vph(X,Z),prog),
    % First verb is -ing form
    { addfeata(Z,ing,Y) }.
```

The predicate VPH_MODAL parses verb phrases introduced by modal auxiliaries. It has two arguments: a structural description of the form vph(Daughters,Features), and a parsing argument which describes the form of the first (modal) verb. VPH_PERF is used to parse the part of the verb phrase after the modal.

```
vph_modal(vph([op([X],[]),hole(Y),word(Z)|W], Nfeats),
    Form) -->
    modal(X,Form), hole(Y), % Modal auxiliary
    { (Type = inf ; Type = base) }, % Base form
```

```
        vph_perf(vph([op([word(Z)],_)|W],Ofeats), Type),
        { addfeata(Ofeats,modal,Nfeats) }.

    vph_modal(vph([op([X],[]),hole(Y),head(Z,W)|V],
        Nfeats), Form) -->
        modal(X,Form), hole(Y), % Modal
        { (Type = inf ; Type = base) },
        vph_perf(vph([head(Z,W)|V],Ofeats), Type),
        { addfeata(Ofeats,modal,Nfeats) }.
        % eg may have insured

    vph_modal(vph([op([X],[]),hole(Y),tail(Z,W)],[tomod,Form]),
        Form) -->
        itomodal(X,Form), hole(Y),
        vph_inf(vph(Z,W)).
        % Modal idiom with to: is to avoid
```

VPH_DO, which has a similar structural description as its only argument, parses verb phrases with *do* periphrasis. It uses VPH_TAIL to parse the verb phrase after the *do* form.

```
    vph_do(vph([op([X],[]),hole(Y)|Z],Feats)) -->
        do(X,Form), { finite(Form) }, hole(Y), % do
        vph_tail(vph(Z,Tfeats),Type,base),
        { addfeata(Tfeats,do,Feats) }.
        % eg did (not) desire
```

Verb phrases containing *have* and *be* as auxiliaries are parsed by the chain of predicates VPH_PERF, VPH_PROG, and VPH_PASS, which call each other successively, after taking care of the respective auxiliaries (if any). These predicates all have two arguments: a structural description and the form of the leading verb.

```
    vph_perf(vph([op([X],[]),hole(Y),word(Z)|W], Nfeats),
        Form) -->
        have(X,Form), hole(Y), % have
        vph_prog(vph([op([word(Z)],_)|W],Ofeats), perf),
        { addfeata(Ofeats,perf,Nfeats) }.

    vph_perf(vph([op([X],[]),hole(Y),head(Z,W)|V], Nfeats),
        Form) -->
        have(X,Form), hole(Y), % have
        vph_prog(vph([head(Z,W)|V],Ofeats), perf),
        { addfeata(Ofeats,perf,Nfeats) }.

    vph_perf(vph(Z, Nfeats), Form) -->
        vph_prog(vph(Z, Nfeats), Form). % No have

    vph_prog(vph([op([X],[]),hole(Y),word(Z)|W], Nfeats),
        Form) -->
        be(X,Form), hole(Y), % be + progressive
        vph_pass(vph([op([word(Z)],_)|W],Ofeats), prog),
        { addfeata(Ofeats,prog,Nfeats) }.
```

```
vph_prog(vph([op([X],[]),hole(Y),head(Z,W)|V], Nfeats),
    Form) -->
    be(X,Form), hole(Y), % be + progressive
    vph_pass(vph([head(Z,W)|V],Ofeats), prog),
    { addfeata(Ofeats,prog,Nfeats) }.

vph_prog(vph(Z, Nfeats), Form) -->
    vph_pass(vph(Z, Nfeats), Form).
    % No be + progressive

vph_pass(vph([op([X],[]),hole(Y)|Z], Nfeats), Form) -->
    be(X,Form), hole(Y), % be + passive
    vph_tail(vph(Z,Ofeats), _, perf),
    { addfeata(Ofeats,pass,Nfeats) }.

vph_pass(vph(X, Nfeats), Form) -->
    vph_tail(vph(X, Nfeats), _, Form).
    % No passive
```

VPH_TAIL parses verb phrases containing a main verb (as described by VPH_MAIN) or a semi-auxiliary (as described by VPH_QUASI). If the main verb is catenative, it must be followed by an infinitive phrase (VPH_INF), possibly with an intervening phrase. The same holds for semi-auxiliaries. VPH_TAIL has three arguments: a structural description of the form vph(Struct,Feats), the type of the central verb (quasi, cat, cop, or main), and the form of the central verb.

```
vph_tail(vph([head([X],[]),hole(Y),compl([Z],
    [])],[quasi]),quasi,Form) -->
    vph_quasi(X,Form),
    hole(Y),
    vph_inf(Z). % With semi-auxiliary

vph_tail(vph([head(X,Y),hole(Z),compl([W],[])],
    Nfeats), cat, Form) -->
    vph_main(vph(X,Y),cat,Form),
    hole(Z),
    vph_inf(W), % With catenative
    { addfeata(Y,cat,Nfeats) }.

vph_tail(vph(X,Y),cop,Form) --> % be as main verb
    vph_main(vph(X,Y),cop,Form).

vph_tail(vph(X,Y),main,Form) --> % Other main verb
    vph_main(vph(X,Y),main,Form).
```

VPH_QUASI parses verb phrases containing a form of *be* and a semi-auxiliary (eg *able*), with up to two intervening phrases. It has two arguments: a structural description of the form vph(Struct,Feats), and the form of the introductory *be*.

```
vph_quasi(vph([cop([X],[]),hole(Y),hole(Z),
    head([word(W)],[])],[quasi]),Form) -->
    be(X,Form),
    hole(Y), hole(Z),
    [W], { has_tag(W,Tag), prefix(Tag,'VXsem') }.
```

VPH_MAIN parses main verbs (as defined by the predicate MAIN_VERB), possibly extended by particles and one intervening phrase. It has three arguments: a structural description of the form vph(Struct,Feats), the type of the verb and the form of the verb.

```
vph_main(vph([head([X],[]),hole(Y),part([Z],[])],[part]),
    Type, Form) -->
    % With particle
    main_verb(X,Type,Form), hole(Y), particle(Z).

vph_main(vph([head([X],[])],[]), Type, Form) -->
    % Without particle
    main_verb(X,Type,Form).
```

VPH_INF, whose argument is a structural description of the form vph(Struct,Feats), parses verb phrases introduced by the infinitive marker and followed by a verb phrase described by VPH_PERF, where the first verb is a base or infinitive form.

```
vph_inf(vph([infmark([To],[]),head(V,Z)],[inf])) -->
    infmark(To),
    { (Type = base ; Type = inf) },
    vph_perf(vph(V,Z),Type).
```

The following three predicates are descriptions of the verbs *do*, *be* and *have* in terms of word class tags:

```
do(word(X),Form) --> [X],
    { has_tag(X,Tag), prefix(Tag,'VD'),
    form(Tag,Form) }.

be(word(X),Form) --> [X],
    { has_tag(X,Tag), prefix(Tag,'VB'),
    form(Tag,Form) }.

have(word(X),Form) --> [X],
    { has_tag(X,Tag), prefix(Tag,'VH'),
    form(Tag,Form) }.
```

The following predicates describe modal verbs and idioms. MODAL includes *had better* (tag VMidi), *need* (tag VMmar), and *may* (tag VM+8), but not *be to* (tag VMito).

```
modal(word(X),Form) --> [X],
    { has_tag(X,Tag), prefix(Tag,'VM'),
    not(prefix(Tag,'VMito')),
    form(Tag,Form) }.
```

ITOMODAL describes modal idioms with *to*: *be (to), have got (to).*

```
itomodal(word(X),Form) --> [X],
    { has_tag(X,Tag), prefix(Tag,'VMito'),
    form(Tag,Form) }.
```

The predicate MAIN_VERB describes main verbs. There are three kinds of main verbs in the sense defined here:

catenative (tag VXcat or VXhav)
copula (tag VB)
main (all others, except modals).

```
main_verb(word(X),cat,Form) --> % Catenative verbs
    [X], { has_tag(X,Tag), ( prefix(Tag,'VXcat') ;
    prefix(Tag,'VXhav') ), puretag(Tag,Ptag),
    form(Ptag,Form) }.

main_verb(word(X),cop,Form) --> % be
    [X], { has_tag(X,Tag), prefix(Tag,'VB'),
    puretag(Tag,Ptag),
    form(Ptag,Form) }.

main_verb(word(X),main,Form) --> % Others
    [X], { has_tag(X,Tag), prefix(Tag,'V'),
    not(prefix(Tag,'VB')),
    not(prefix(Tag,'VX')), not(prefix(Tag,'VM')),
    puretag(Tag,Ptag), form(Ptag,Form) }.
```

The infinitive marker is characterized by its tag TO:

```
infmark(word(To)) --> [To], {has_tag(To,'TO')}.
```

Particles are the following words, when they are tagged AAspa on the word class level:

```
particle(word(X)) -->
    [X], { has_tag(X,Tag), prefix(Tag,'AAspa'),
    word(X,W), ( W = in ;  W = about ; W = on ;
    W = up ; W = out ; W = down ;  W = back ;
    W = off ; W = over ; W = away ; W = along ;
    W = through ; W = round ; W = after ;
    W = around ; W = to ; W = together ) }.
```

The predicate FORM characterizes the form of a verb as one of the alternatives base, pres, inf (only *be*), past, perf, and prog.

```
form(Ptag,base) :- suffix(Ptag,'+0').
form(Ptag,base) :- suffix(Ptag,'+0DN').
form(Ptag,base) :- suffix(Ptag,'+0N').

form(Ptag,pres) :- prefix(Ptag,'VM+'),
    suffix(Ptag,'+8').
form(Ptag,pres) :- suffix(Ptag,'+3').

form('VB+0',inf). % be
form('VB+4',pres). % are
form('VB+5',past). % was
form('VB+6',past). % were

form(Ptag,past) :- prefix(Ptag,'VM+'),
    suffix(Ptag,'+9').
form(Ptag,past) :- suffix(Ptag,'+D').
form(Ptag,past) :- suffix(Ptag,'+DN').
form(Ptag,past) :- suffix(Ptag,'+0DN').

form(Ptag,perf) :- suffix(Ptag,'+N').
form(Ptag,perf) :- suffix(Ptag,'+DN').
form(Ptag,perf) :- suffix(Ptag,'+0N').
form(Ptag,perf) :- suffix(Ptag,'+0DN').

form(Ptag,prog) :- suffix(Ptag,'+G').
```

The next predicate characterizes finite verb forms:

```
finite(base).
finite(pres).
finite(past).
```

The next two predicates describe participles:

```
psp(W) :- % Present participle
    has_tag(W,Tag), puretag(Tag,Ptag),
    form(Ptag,prog).

ptp(W) :- % Past participle
    has_tag(W,Tag), puretag(Tag,Ptag),
    form(Ptag,perf).
```

'Holes' in verb phrases may be empty, or occupied by noun phrases or adverb phrases:

```
hole(X) --> nph(X). % Noun phrase
hole(X) --> aph(X,_,_). % Adverb phrase
hole(nil) --> []. % Empty
```

### 4.3.9 Remaining input

Input words that cannot be parsed as phrases of the above types, are taken care of by the predicate OTHER, which simply makes them elements of structures headed by the functor OTHER:

```
other(other(X)) --> [X].
    % Punctuation marks, conjunctions, ...
```

### 4.3.10 Auxiliary predicates for data access etc

```
% Handling lists of features:

% Check presence of features in list:

% Is feature Feat in feature list Featlist:
infeatlist(Feat,Featlist) :-
    member(Feat, Featlist).

% Check that none of the features in the first
% list are present in the second list:
nofeats([],_).
nofeats([Feat|Rest],Featlist) :-
    not(infeatlist(Feat,Featlist)),
    nofeats(Rest,Featlist).

% Check that all of the features in the first list
% are present in the second list:
allfeats([],_).
allfeats([Feat|Rest],Featlist) :-
    infeatlist(Feat,Featlist),
    allfeats(Rest,Featlist).

% Check if any of the features in the first list
% is present in the second list:
anyfeat([Feat|Rest],Featlist) :-
    infeatlist(Feat,Featlist), !.
anyfeat([_|Rest],Featlist) :-
    anyfeat(Rest,Featlist).

% Add features to list, suppressing double entries:

% Insert as first element:
addfeata(Oldlist,Feat,Oldlist) :-
    infeatlist(Feat,Oldlist), !.
addfeata(Oldlist,Feat,[Feat|Oldlist]).

% Insert as last element:
addfeatz(Oldlist,Feat,Oldlist) :-
    infeatlist(Feat,Oldlist), !.
addfeatz(Oldlist,Feat,Newlist) :-
    append(Oldlist,[Feat],Newlist).
```

The parser also uses the following predicates, whose definitions will not be specified in detail:

Access to structural descriptions:

| | |
|---|---|
| GET_HEAD_WORD | find head word in structure |

Access to tagged Brown corpus input in 'vertical' format:

| | |
|---|---|
| HAS_TAG | access to full word class tag (including length) |
| PURETAG | access to word class tag proper (not including length) |
| WORD | access to word body |

String handling:

| | |
|---|---|
| PREFIX | test for prefix |
| SUFFIX | test for suffix |

List handling: (standard)

| | |
|---|---|
| APPEND | list concatenation |
| MEMBER | list membership |

5

# Lexical items peculiar to spoken discourse

*Anna-Brita Stenström*

## 5.1 Is there a special spoken lexis?

Some lexical items are much more common in speech than in writing. A list comparing the most frequent words in the London-Lund Corpus (LLC) with the most frequent words in the Lancaster-Oslo/Bergen Corpus (LOB) (p 17) shows for instance that the verbs *know*, *think*, and *mean* with rank 15, 25, and 44 in LLC, do not appear at all in the LOB list; nor does the adverb *really*, ranked 59 in LLC, while *well* ranked 14 in LLC appears only at rank 95 in LOB. Not surprisingly, items like *yes*, *no*, and *oh* do not figure among the most frequent words in the written corpus. Such lexical items are typical of interactive speech and closely bound to the communicative situation; others like *really* and *well* abound in spoken discourse for other reasons.

Many lexical items which are typically used in spoken discourse (henceforth referred to as 'D-items') are awkward or even impossible to analyse in syntactic terms, and they generally contribute little, if anything, to the propositional content of the utterance. However, the mere fact that a lexical item is extremely common in spoken discourse does not immediately qualify it as a D-item. The decisive factor is whether its interactive and pragmatic functions prevail over its 'ordinary' grammatical function. This is illustrated in (1):

(1) but I ||don't 'really KNOW that 'I'm■ ||going to 'be a 'vast a'mount of HELP to 'you■ • I was || INTERESTED in your {AD||VERTISEMENT■ }■ and ||and [əm] -- • [? ə] but [ə] I ||gather you're ▵AFTER■ «an» e||normous a'mount of 'INFOR▵MATION■ and I ||don't REALLY 'know that I've 'got - ||YOU 'know■ whether ||WHAT I've 'got■ ||is [?] of ▵any ▻HELP■ I mean it's ||really for 'you to DECIDE■ ||REALLY■ --- (S.2.2:16-25)

None of the four instances of *really* will be regarded as a D-item since they all retain too much of the adverbial function. What finally determines what the item does in the discourse is its position in the complete contextual situation; it may serve as a syntactic constituent in a sentence or as a move in the interaction.

*Really*, which will be discussed in detail in Section 5.5, can be used as a prototype to illustrate how the same lexical item can serve more than one function, depending on where it occurs. In syntactic terms, it may serve as an emphasizer, adding to the force of an adjective (cf Quirk et al 1985:447), as in (2) where it modifies the adjective *annoying*:

(2) [? i] it's ||really ANNOYING■ (S.2.13:82)

or as a content disjunct (Quirk et al 1985:620 ff), reflecting the speaker's attitude to what he is saying and modifying the whole utterance:

(3) ||didn't make any ▵DIFFERENCE■ ||REALLY■ (S.1.1:1050-1051)

In interactional terms, *really* may serve as a 'react' showing the speaker's attitude to the previous speaker's utterance; as a 'go-on', passing on the turn at a transition relevance point (Sacks et al 1974) and encouraging the current speaker to continue (cf 'continuer' in Schegloff 1982); as a 'follow-up' in a question-response exchange, constituting the questioner's confirmation of the response; as a 're-opener' querying the response; or as a 'planner', with concurring syntactic and interactive functions (see further pp 152ff).

The term 'D-item' will be used to designate lexical items that occur almost exclusively in spoken interaction, for example *yes*, *please*, *sure*, *shut up*, and question tags (Q-tags), as in (4):

(4) A: ... but I ‖thought I △WOULDN'T 'have an ABSTRACT■ be‖cause I △think you △have to [ə] • sup△ply ▻something YOUR△SELF to the ▻picture■ ‖HAVEN'T you■

B: ‖SURE■ (S.1.8:479-482)

The term also designates lexical items that occur in both written and spoken discourse but with a particular function in speech, for example *sort of*, *you know*, and *well*:

(5) whereas ‖HART■ I ‖mean as you KNOW■ sort of -- (S.1.5:622-624)

as well as lexical items that are particularly frequent in spoken discourse but would have a similar discourse function in speech and writing, eg *anyway* and *now* (as a transitional device):

(6) *but* ‖that didn't △HAPPEN■ until ‖LONG 'after [ði:]■ - • [ə:] -- ‖British and 'French and A△merican - ARMIES■ had ‖really sort of --- ‖anyway I'm △SORRY■ I was DI‖GRESSING■ • but ‖what I 'mean IS■ -- (S.2.3:362-367)

Summing up, D-items may be realized by single words like *well*, *oh*, *all right*, and *anyway* or by longer strings like *as you know* and *I'm sure that's right*, and they are used for taking, keeping, and yielding the turn by performing a speech action, for empathizing with the listener, or for organizing the message.

## 5.2 The structure of spoken discourse

Spoken interaction will be viewed in terms of four hierarchical levels: exchange, turn, move, and act.

The **exchange** is the minimal interactive unit which consists of at least two consecutive turns (and at least two moves) produced by different speakers.

The **turn** is what each speaker says before the next speaker takes over; it consists of one or more moves. Turns can be opening, holding-up, releasing, continuing, and terminating.

The **move** is what the speaker does in order to start, carry on, and finish an exchange, eg Initiate-Respond. It consists of one or more acts. A simple move consists of one act, eg a response proper; a compound move consists of more than one act, eg a response proper followed by an 'emphasizer'.

The **act** involves the illocutionary and pragmatic content, eg request-accept.

This model is slightly different from that of Sinclair & Coulthard (1975) in that I have introduced the level of 'turn' to handle utterances containing more than one 'move'. I distinguish the following moves and acts, most of which will be illustrated in the following discussion (see further Stenström 1984a, 1984b):

MOVES

**Call-off** is the first part of a pre-closing or closing exchange.

**Close** is the second part of a pre-closing or closing exchange.

**Follow-up** terminates an exchange and involves speaker-shift.

**Frame** introduces new topics (aspects) and new exchanges.

**Go-on** indicates that the listener is paying attention to what is being said and encourages the current speaker to go on.

**Initiate** initiates an exchange by inquiring, informing, suggesting, etc.

**Question** elicits a response.

**Re-open** queries what was stated in a response or an inform.

**Response** answers a question or supports or challenges the previous speaker's initiating or re-opening move.

ACTS

**Accept** accepts a request, a suggestion or an opinion.

**Apology** serves to apologize.

**Clarify** disambiguates a previous utterance by the same speaker.

**Comment** adds information not expressed in a previous act in the same move.

**Conclude** draws a conclusion from a previous utterance.

**Confirm** responds to a request for confirmation.

**Direct** orders the addressee to do something.

**Elicit** is used as an umbrella term for any type of question.

**Emphasizer** highlights a preceding act in the same response, go-on, or follow-up move.

**Endorse** supports the point made by the preceding speaker.

**Evaluate** expresses the speaker's attitude.

**Exemplifier** introduces more detailed information.

**Expletive** expresses annoyance, excitement, pain, etc.

**Filler** serves as a verbal pause.

**Frame** marks a boundary in the discourse, indicates that the speaker introduces a new topic/aspect or sums up what has already been said.

**Greet** is what you do when you meet or leave somebody.

**Hedge** reflects uncertainty and vagueness, avoids abruptness.

**Inform** supplies 'neutral' information which does not explicitly reflect the speaker's attitude.

**Object** objects to the previous speaker's utterance.

**Planner** serves as a temporary substitute for a clause element.
**Please** emphasizes the speaker's wish and marks politeness.
**Prompter** appeals for feedback.
**React** shows the speaker's attitude to a previous request, asks the listener to do something.
**Smooth-over** responds to an apology.
**Softener** empathizes with the listener.
**Staller** is used to gain time.
**Suggest** comes up with a proposal.
**Thanks** expresses gratitude.
**Uptake** connects the previous speaker's last move with the succeeding move uttered by the speaker who produced the uptake.

The most convenient starting-point for identifying what speech signals do in a dialogue is the conversational exchange and the speaker turns. D-items can occur in more than one position in the exchange structure and the turn. The items generally perform different functions in different positions but can also perform different functions in the same position. Naturally, the function of an item does not depend only on its position but also on its own inherent meaning and the larger context. Conversational exchanges vary in size and complexity. In this corpus, two-turn exchanges were somewhat more common than three-turn exchanges, but exchanges consisting of up to five and six turns were not unusual.

If the discourse signal makes up a turn of its own, it is also a move consisting of one act. Otherwise it realizes either a move within the turn or an act in a move. Its pragmatic function varies with its vertical (turn-by-turn) position in the exchange as well as with its horizontal (within-the-turn) position (see pp 164ff). Turns have the following functions:

| | |
|---|---|
| Turn [1] | **A**: OPEN |
| Turn [2] | **B**: CONTINUE/TERMINATE |
| Turn [3] | **A**: HOLD-UP/TERMINATE |
| Turn [4] | **B**: TERMINATE |

*Right* in (7) terminates the two-turn exchange by responding to an informing initiating move which opens the exchange. In the following three-turn-exchange (8), *right* is part of another response move, one that continues an exchange terminated by a follow-up move. The next two examples illustrate four-turn exchanges; *right* in (9) holds up the exchange by a re-opening move, and *right* in (10) terminates the exchange by a confirming follow-up move.

| **Example** | | | **TURN** | **Move** | **act** |
|---|---|---|---|---|---|
| (7) Two-turn exchange (S.8.2:1009-1010) | | | | | |
| [1] | A: | it's UNDER■ – ‖H for HARRY■ | OPEN | Initiate | inform |
| [2] | B: | ‖RIGHT■ | TERMINATE | Respond | accept |
| (8) Three-turn exchange (S.8.1:822-827) | | | | | |
| [1] | A: | so he ‖knows ˈwhat it's ABOUT■ | OPEN | Initiate | conclude |
| [2] | B: | ‖RIGHT■ ‖OK■ ∗•MHM■ ∗ | CONTINUE | Respond | confirm |
| [3] | A: | ∗‖GOOD■ ‖THANK you■ ∗ | TERMINATE | Follow-up | endorse |
| (9) Four-turn exchange (S.9.1:312-315) | | | | | |
| [1] | A: | [ə:m] ‖shall we say • would ‖twelve o'Δclock be OKAY■ | OPEN | Initiate | suggest |
| [2] | B: | ‖LOVELY■ | CONTINUE | Respond | accept |
| [3] | A: | ‖RIGHT■ | HOLD-UP | Re-open | elicit |
| [4] | B: | ‖YES■ | TERMINATE | Respond | clarify |
| (10) Four-turn exchange (S.7.2:114-118) | | | | | |
| [1] | A: | so ‖what ˈtime are you COMING this ˈafternoon■ | OPEN | Initiate | elicit |
| [2] | B: | [əm • əmə] ‖as we Δsaid about ▵four O'CLOCK■ | CONTINUE | Respond | inform |
| [3] | A: | ‖OK■ ‖YEAH■ | TERMINATE | Follow-up | confirm |
| [4] | B: | ‖RIGHT■ | TERMINATE | Follow-up | confirm |

The fact that the function of a discourse item is related not only to its position in the exchange, as we saw in the previous examples, but also to its position in the turn will be discussed in Section 5.6.

## 5.3 The classification of D-items

The discourse level was introduced in our tagging-system for the analysis of words and expressions that were found to serve interactive and pragmatic rather than grammatical functions. We felt that these D-items could not be adequately accounted for at the word, phrase and clause levels of analysis in a three-level tagging-system. At the discourse level, items were analysed in terms of speech-organizing, interactional, and communicative devices. Both tagging-systems are illustrated in the analysis of (11):

(11) >A: • ∆mid ↗APRIL■ • ||we had ∆reached the →POINT■ of ||thinking that we ∆weren't 'going to be 'able to ∆↗REACH■ • a ||↗POLICY de▻cision– *•* and «so» ||we

B: *||that's ↘RIGHT■ *

>A: must • [?] ∆tell these ↗GUYS■ *«that we'll» carry ↘ON■ –* (S.1.2:165-172)

*That's right* can be analysed in strict syntactic terms (cf pp 96ff for the tag labels):

| | | | |
|---|---|---|---|
| TEXT | that's | | right |
| WORD | RD*VB+3 | | JA |
| PHRASE | NPH:dem | VPH:pres | JPH |
| CLAUSE | S | V | C |

The following is an alternative analysis in discourse terms:

| | | |
|---|---|---|
| TEXT | \|\|that's | ↘RIGHT■ |
| WORD | DR2 | |
| PHRASE | 0 | 0 |
| CLAUSE | 0 | 0 |
| DISCOURSE | RESP | |

In the second analysis, the clause *that's right* is considered to be one D-item, hence at word class level labelled DR2 denoting a D-item serving as a response signal (R) and consisting of two words (2). It is left untagged at the phrase and clause levels; at the discourse level it is tagged RESP for its interactive function of response in the discourse.

Yet, this second analysis is not entirely satisfactory. Unlike single words like *yes* and *OK*, the string *that's right* can also be analysed syntactically (as in our first analysis), and if we refrain from this option altogether, useful generalizations will be lost.

The classification of D-items was based on the study of ten LLC texts, most of them face-to-face, two-party, unplanned dialogues. The following exemplified discourse categories were identified (DA, DB, etc are the tag codes) in the sample:

| | | |
|---|---|---|
| Apologies (DA) | DA | *pardon, sorry* |
| | DA2 | *excuse me, I'm sorry* |
| | DA4 | *I beg your pardon* |
| Smooth-overs (DB) | DB2 | *don't worry, never mind* |
| Hedges (DC) | DC2 | *kind of, sort of* |
| | DC3 | *sort of thing* |
| Expletives (DE) | DE | *damn, gosh, hell* |
| | DE2 | *fuck off, good heavens, the hell* |
| | DE3 | *for goodness sake, good heavens above, oh bloody hell* |
| Greetings (DG) | DG | *hi, hello* |
| | DG2 | *good evening, good morning* |
| | DG3 | *Happy New Year, how are you* |
| | DG4 | *how do you do* |
| Initiators (DI) | DI | *anyway, however, now* |
| Negative (DN) | DN | *no* |
| Orders (DO) | DO2 | *give over, go on, shut up* |
| Politeness markers (DP) | DP | *please* |
| Q-tags (DQ) | DQ2 | *is it* |
| | DQ3 | *isn't it* |
| Responses (DR) | DR | *ah, fine, good, uhuh, oh, OK, quite, really, right, sure* |
| | DR2 | *all right, fair enough, I'm sure, I see, that's good, that's it, that's right, that's true, very good* |
| Softeners (DS) | DS2 | *I mean, mind you, you know, you see* |
| | DS3 | *as you know, do you see* |
| Thanks (DT) | DT | *thanks* |
| | DT2 | *thank you* |
| *Well* (DW) | DW | *well* |
| Exemplifiers (DX) | DX | *say* |
| Positive (DY) | DY | *mhm, yeah, yes, yup* |

The D-categories presented in the list fall into three groups depending on their tendency to constitute a separate move.

Group (*a*): categories which do not constitute a separate move;
Group (*b*): categories which may or may not constitute a separate move;
Group (*c*): categories which generally constitute a separate move.

**Group (*a*)**
Categories which do not constitute a separate move include hedges, initiators, softeners, exemplifiers and *well*. The items in this group share certain general features. They realize acts in various types of move:

(12) *well* I don't know
(13) she *sort of* said to me that ...

They are not meaningful or informative in isolation but may add some dimension to the propositional content of the move, without altering its function (with the exception of exemplifiers and initiators):

(14) *now* all this happened last night
(15) *well* yes

They are all turnholding, indicating that there is more to come (with the exception of turn-final softeners):

(16) last night *you know* (something extraordinary happened)

Whether an item is turntaking, turnholding or turnyielding is of course related to its position. Only softeners fulfil all three functions:

(17) *you know* he said to me ...
he said to me *you know* that ...
that's what he said to me *you know*

Softeners act differently depending on whether they occur in initial, medial, or final position (cf Crystal & Davy 1975:92-97). Softeners in final position are unique in that they serve as explicit appeals for feedback, especially if they carry a rising tone. They are therefore typically turnyielding and interpersonally oriented. By adding a softener or a Q-tag, the speaker shows that he is not completely certain of what he is stating or, alternatively, he pretends not to be but leaves the final decision to the listener (cf Östman 1981). In other words, he adopts a face-saving strategy involving politeness. Q-tags in final position have a stronger elicitative force than softeners and are generally followed by a confirming reaction regardless of pitch contour (cf Stenström 1984a:224).

Initiators and exemplifiers are purely organizational devices providing very little information. Initiators indicate that something new is to come, eg a new aspect of the topic already being discussed, or they direct the conversation back to a previous stage in the discourse after a momentary digression. Exemplifiers introduce illustrations and details.

### Group (*b*)

Categories which may or may not constitute a separate move include politeness markers (*please*), comment Qs (*isn't it*, *does it*, etc), *no* and *yes*. The items in this group may make up the speaker's whole contribution (turn). If they do, they constitute a turn consisting of one move. Otherwise they are acts in moves:

| | | |
|---|---|---|
| (18) | A: can I borrow your car? | |
| | B: *please* | **one turn** = one move consisting of one act |
| | A: *please* don't go | **one turn** = one move consisting of two acts |

Similarly:

| | | |
|---|---|---|
| (19) | A: has Bryan arrived? | |
| | B: *yes* | **one turn** = one move consisting of one act |
| | A: has Jean come too? | |
| | B: *yes* she has | **one turn** = one move consisting of two acts |

Response items realizing acts in a move can be found in initial, medial and final position:

(20) *yes* that's right
that's right *yes*
he did *yes* that's right

Items from all the categories can occur in response position and serve as a response (if taken in the broad sense of 'reaction'):

| | | |
|---|---|---|
| (21) | A: can I sit down | |
| | B: please | Politeness |
| (22) | A: I heard him saying it | |
| | B: did you | Comment Q |

The items in this group are inherently turntaking. Only *please* may be explicitly turnyielding:

(23) A: is this the English department please
B: certainly

Notice the difference between comment Qs and Q-tags: comment Qs can be used as responses (feedback) whereas Q-tags are used as response-inviters:

| | | |
|---|---|---|
| (24) | A: this isn't the way he did it | Inform |
| | B: isn't it (= I see) | Response (feedback) |
| (25) | A: this I think is looking at the problem from a completely different view isn't it (= don't you think) | Elicit |
| | B: yes | Response (confirm) |

**Group (*c*)**

Categories which generally constitute a separate move include: apologies (*sorry*), smooth-overs (*never mind*), expletives (*hell, fuck off*), greetings (*hi*), orders (*shut up*), responses (*mhm*, *sure*, *that's right*, *really*), thanks (*thank you*). These items are inherently turntaking or turnyielding. Yet greetings, which belong to so-called 'adjacency-pairs', are both; the first greeting triggers off another greeting in reply:

(26) A: hello
B: hello

The items in this group also denote a strong interpersonal relationship:

| | | |
|---|---|---|
| (27) | A: I'm sorry | Apology |
| | B: never mind | Smooth-over |

The category of responses (DR) is mixed and includes items that realize different moves in different positions, ie depending on what type of move precedes or follows. This category must therefore be subcategorized before it can be satisfactorily dealt with. 'Response' is here taken in the very broad sense of 'reaction to any utterance made by A' (greetings and thanks excluded):

| | | |
|---|---|---|
| (28) | A: he said he'd do it | |
| | B: *fine* | Go-on |
| | > A: at once | |
| (29) | A: he didn't do it | |
| | B: *I see* | Follow-up |
| (30) | A: did he do it | |
| | B: *sure* | Response |

(31) A: he did it
B: *really* Re-opener
A: yes indeed

Four categories - comment QS, *please*, responses, and greetings - belong to different groups but yet have an important feature in common: they occur both as eliciting/inviting devices and as elicited/invited elements.

The interpersonal relationship, the most characteristic feature of conversation, can be observed in all the categories. In group (*a*) only softeners are explicitly interpersonal (A/B-oriented). The other *a*-categories are chiefly message-oriented discourse organizers. All the categories in groups (*b*) and (*c*) are interactive and therefore interpersonal.

## 5.4 A monologue and a dialogue compared

Assuming that the use of discourse items would differ in dialogue and monologue, I selected two LLC texts for a comparative study: the dialogue S.4.1, in which a married couple are chatting over lunch, and the monologue S.12.6, in which a former master builder talks about memories from his childhood. Both texts were non-surreptitiously recorded. This probably did not affect the former master builder who gave a lecture to an audience, but it obviously had an effect on the young couple. First, they refer in their conversation to the fact that it was being recorded and that their speech was somewhat strained; second, they also bring up a large number of topics for discussion, which obviously indicates their awareness that they were supposed to go on speaking for a certain period of time and hence felt forced to start on a new topic as soon as the previous one was exhausted.

All categories of D-items were much less frequent in the monologue than in the dialogue, which was to be expected. In the monologue, most of the categories belonging to groups (*b*) and (*c*) were not expected to occur at all, possibly with the exception of expletives and apologies. On the other hand, items from group (*a*), which reflect the planning process, might occur in either type of spoken discourse.

Four important factors play a crucial role for the difference in use between the two types of talk. First, the monologue was preplanned - the speaker had worked out beforehand what he was going to say - whereas the dialogue was spontaneous. Second, the speaking situations were different. The speaker in the monologue could go on speaking without risking interruption, while the parties in the dialogue had to take turns. This required cooperation. Third, the speaker's strategy in a the dialogue was affected by the hearer's reactions in the

**Table 5:1.** D-categories in the dialogue and the monologue.

| ITEM | DIA | MONO | TOTAL |
|---|---|---|---|
| ORGANIZING | | | |
| *Well* (DW) | 37 | 10 | 47 |
| Softeners (DS) | 33 | 4 | 37 |
| Expletives (DE) | 13 | 1 | 14 |
| Initiators (DI) | 3 | 7 | 10 |
| Hedges (DC) | 7 | - | 7 |
| Exemplifiers (DX) | 1 | - | 1 |
| | | | |
| INTERPERSONAL | | | |
| Positive (DY) | 97 | - | 97 |
| Negative (DN) | 28 | - | 28 |
| Responses (DR) | 20 | - | 20 |
| Q-tags (DQ) | 16 | - | 16 |
| Apologies (DA) | 4 | - | 4 |
| Orders (DO) | 4 | - | 4 |
| Politeness markers (DP) | 2 | - | 2 |
| Thanks (DT) | 1 | - | 1 |
| Total | 266 | 22 | 288 |

form of oral and/or visual feedback signals, which means that he might, for example, have to start replanning in the middle of his performance in order to be understood. Fourth, the monologue was public, and the dialogue was private.

A speech-situation in which the parties are unprepared, but aware that they are being recorded and should not stop speaking, certainly invites hesitation phenomena, such as pauses (filled and unfilled), verbal fillers, and restarts. Such hesitation phenomena usually occur at the beginning of utterances when the speaker has not yet made up his mind how to continue (cf Brown 1977:120-24). As will be seen in Chapter 8, clusterings of such items were common.

Table 5:1 shows the frequency and distribution of D-categories in the dialogue and the monologue with the items arranged in frequency order. The different speech-situations are very clearly reflected in the use of D-items. Practically all of the 16 categories (p 144) were found in the dialogue but only four in the monologue. As can be expected, interpersonal D-items occurred only in the dialogue. Note especially the large number of response-items (DY, DN and DR). The D-categories found in the monologue belong to the area of planning and organizing.

Softeners were realized somewhat differently in the two text types: by *you know* in the dialogue and by *as you know* in the monologue with a slight shift of meaning with consequences for the A/B orientation. *As you know* signals that the speaker presupposes that the listeners know what he is referring to, and therefore acts as a politeness device without appealing for verbal feedback. *You know* in final position (and with a rising tone) does appeal for feedback, but without necessarily presupposing previous knowledge.

Hedges, here represented by *sort of*, did not occur at all in the monologue, probably a consequence of the pre-planning. It seems, however, that hedging is not only situation-bound but also highly idiosyncratic. Similarly, the use of Q-tags and softeners is largely speaker-specific.

According to Table 5:1, initiators were more common in the monologue than the dialogue. This is not the whole truth, however, since *well* - in a category of its own - served as an initiator in eight out of the ten cases in the monologue. This makes initiators the typical D-category. Note that an initiator is not just an item that occurs frequently in initial position, a definition that would have included the majority of all instances of *well*, *yes* and *no* in the dialogue. Initiators do very specific things besides occurring initially. This is illustrated in (32)-(34):

(32) HOW‖ÉVER■ • my ‖mother 'played the PIÁNO■ (S.12.6:233-234)

(33) ‖NOW■ – ‖after the ALLÚYSONS■ • [ə:] Sir ‖Philip WÌLD■ a di‖rector of the ▵Port of 'London AUTHÓRITIE«S»■ • ‖stopped ••• (S.12.6:733-737)

(34) ‖WELL■ AN‖ÒTHER 'great 'family■ ‖lived HÉRE – {‖during my LÍFE■ }■ (S.12.6:715-717)

*However* in (32) resumes the narrative after a short digression; *now* in (33) points forward and introduces new information ('let's proceed'); *well* in (34) links one part of the narrative to what follows ('let's leave this and go on'). Svartvik (1980a) uses the term 'qualifier' for *well* serving as a link between previous and following discourses. Sinclair & Coulthard (1975) call it a 'marker' which realizes a 'Framing' move. In later sections I shall use the term 'frame' instead of 'initiator'.

*Well* in the dialogue acted differently depending on whether it introduced an elicit, a response or an inform (taken in a broad sense and including various types of comments and retorts). Compare (35)-(37):

(35) a: well ||can you can you 'not get CLOSER■

b: ||yes I ▵COULD have DONE■ (S.4.1:81-82)

(36) a: well ||what are we 'doing 'this WEEKEND■ •••

b: well I've ||nothing DOWN ▷anyway • {at ||ALL■ }■ (S.4.1:14-17)

(37) b: when I ||READ it■ I ||meant to 'point it △OUT to you■

a: ||WELL■ it ||wasn't a bad SU△GGESTION■ ||REALLY■ (S.4.1:801-805)

In (35) *well* links a question to a previous utterance in the dialogue, at the same time introducing a new aspect. In (36) a new topic is brought up but there is no link with what preceded - *well* is equivalent to *now* in the same position. The second *well* in (36) serves as a response-prefix, typically introducing a response which is insufficient in some respect; B cannot come up with a good suggestion and therefore does not produce a direct answer but a response which lets A understand, implicitly, that he has no plans for the weekend instead of saying so straight out. In (37), finally, *well* introduces a move expressing evaluation and acts very much as a softener. Broadly speaking, *well* serves as an initiator with questions, as a prefix indicating indirectness and insufficiency with responses, and as a softener with informs.

*Yes* and *no* have been provided with individual tags because of their high frequency in speech (instead of being included in the general DR category). Although their main function is to serve as polar responses, they are also met with as go-ons and follow-ups. The go-on function was not very prominent in this dialogue, however, probably because the parties knew each other so well that feedback in that form was not necessary.

The go-on function is also met with in comment QS which otherwise behave similarly to *really* and also serve as follow-ups and re-openers:

(38) A: the weather isn't going to be better

B: isn't it — Go-on

> A: until next week

(39) A: the weather isn't going to be better

B: isn't it — Follow-up

(40) A: the weather isn't going to be better
B: isn't it Re-opener
A: no I heard the weather forecast

Summing up, the discourse (D) level was adopted as part of our analysis to handle speech-specific items that cannot be appropriately taken care of at the syntactic level. In this study based on a small sample, it appeared that D-items of the types recognized here were generally much less common in the monologue than the dialogue and, furthermore, that fewer D-types were used in the monologue. This is basically a consequence of the disparate speech-situations, but it probably also reflects individual differences in speech-behaviour. As can be expected, genuine interpersonal items were altogether absent in the monologue. But D-items classified as 'organizing' were also less frequent, one obvious reason being that the monologue had been prepared in advance. To what extent the use of D-items is related to the monologue/dialogue situation as such, or is speaker-specific, cannot be stated on the basis of only two text samples, but I presume that an unplanned monologue in a less formal situation would contain a much greater number of D-items from a larger variety of categories.

Since dialogue is inherently interactive, the devices of turntaking, turnkeeping, and turnyielding are crucial for a smooth conversation. Communicative cues such as softeners do not only reflect the speaker's personal involvement but also add liveliness to the conversation. If such signals were altogether lacking, the speakers might sound not only uninterested but also plain boring. The same applies to monologues: a narrator or a lecturer who drops all communicative cues runs a considerable risk of losing the attention of his audience.

## 5.5 *Really*

Among the various D-items I made a special study of *really* and the '*right* set' (see pp 161ff). One of the reasons for studying *really* was its frequency in the spoken material, another its obvious functional variety which seemed to be somewhat different in speech and writing. My observations are based on a sample of LLC (from the LLC:o version) consisting of approximately 170,000 words for the spoken material and on the entire LOB corpus of approximately one million words for the written data. I found that *really* occurred with a density of 3.17 per 1,000 words in LLC and 0.31 per 1,000 words in LOB, a considerable difference which marks *really* as a characteristic feature of conversation. To facilitate the comparison I picked a random sample consisting of

**Figure 5:1.** Functions of *really*.

| | | |
|---|---|---|
| Intensifier: | | she's really nice |
| Evaluater: | | REALLY■ I do■ |
| Re-opener: | A: | this is what I HEARD■ |
| | B: | REALLY■ |
| | A: | YES■ |
| Go-on: | A: | this is what I HEARD■ |
| | B: | REALLY■ |
| | > A: | right NOW■ |
| Planner: | | he's a really he's a THOROUGHBRED |

100 instances of *really* from each concordance. I identified five functions of *really* (see Figure 5:1):

As an **intensifier**, integrated in the clause structure and placed next to a head, *really* is part of a syntactic unit.

As an **evaluater** it is peripheral to the clause structure and reflects the speaker's attitude to the entire predication.

As a **re-opener** expressing the speaker's reaction to what the previous speaker said and eliciting a response, it is a purely interactional device.

As a **go-on** signalling that the hearer is an active listener and expecting the current speaker to go on talking, it is also purely interactional.

As a **planner** making it easier for the speaker to formulate his message by providing an opportunity for reconsideration, it is partly integrated in the clause structure.

There is a clear relationship between the function and the position of D-items. Quirk et al (1985:583-584) remark that 'emphasizers', placed next to a particular element and not separated by intonation or punctuation, often emphasize that element alone, but that there may be ambivalence as to whether the emphasis is on the part (a single constituent) or on the whole (the proposition). The problem may be highlighted in the following way:

(41a) this question is *really* surprising
(41b) this is a *really* surprising question
(41c) this is *really* a surprising question
(41d) this *really* is a surprising question

(41e) *really* this is a surprising question

(41f) this is a surprising question *really*

Disregarding the possible effect of prosody or punctuation, it may safely be stated that *really* placed next to the adjective *surprising* is clearly an intensifier. But the further it is moved towards the left, the less emphasis there is on the part (*surprising*), and the more there is on the whole (*this is a surprising question*). In (41e), where *really* is placed initially, it no longer intensifies a single clause element but comments on the entire proposition. But how are we to describe the use of *really* in clause-final position (41f)? Greenbaum (1969:144) states that *really* is unambiguously a disjunct in initial position and also usually when it occurs in a separate tone unit; it 'makes explicit the speaker's view that the statement being made is true'.

*Really* in initial and medial position (41a-e) has already been dealt with in detail by eg Greenbaum (1969), Bolinger (1972a), Jacobson (1978), and Quirk et al (1985); but, to my knowledge, very little has been said about *really* in final position (41f), where its function seems to be particularly doubtful and where it was highly frequent in the present data.

There was a clear tendency for *really* to collocate with negation, usually realized by *not*; one third of all instances of *really* appeared in a negative surrounding in the conversations and one fourth in the written texts. The most frequent position of *really* in negative declaratives is illustrated in (42) and (43) with *really* within the scope of clause negation (Quirk et al 1985:587):

(42) and I ‖haven't really ▵CÒNCENTRATED■ (S.3.1:172)

(43) This place isn't really a Political Centre (M02 55)

The syntactic configurations in which *really* occurred in positive declaratives were much more complex and varied. But the dominating position was immediately after the operator or BE, ie the same as in negative declaratives:

(44) I have really got a bug about it (W.1)

(45) Some of the hill-climbs, too, were really devastating (G24 157)

As mentioned above, clause-final position of *really* was particularly common in the spoken data (17% in speech vs. 4% in writing):

(46) and ‖Eileen has FOR▵GÒTTEN■ that ‖RÈALLY «you ▷know»■ • (S.1.13:462-463)

(47) it's ‖all △M↘AD R↗EALLY■ -- (S.1.10:1266)

Note, however, that *really* very seldom occurred in clause-final position in negative utterances but was usually placed immediately after the negation. In interrogatives, the position of *really* was not directly comparable in the two media due to the different realizations of the interrogative form, with inverted word-order in writing (48-49) and tag-questions in speech (50-51):

(48) Don't you really know?

(49) Are you really happy with him?

(50) goes 'on ⊳really R∨EGULARLY 'does it■ (S.3.3:810)

(51) it ‖was △really CONV↘INCING■ ‖W↘ASN'T it■ (S.2.10:839-840)

End of clause correlated with end of tone unit in the spoken data where *really* was preceded by a tone unit boundary in 34% of all clause-final occurrences. In LOB it was occasionally separated from the rest of the clause by a comma:

(52) (...) put the △weakest candidate F→IRST■ ‖R↗EALLY■ (S.1.3:490-491)

(53) She lost her way, really (P06 143)

The comma in writing often corresponded to a pause in the spoken data:

(54) be‖cause we △don't know 'what it M↗EANS■ - ‖R↗EALLY■ (S.3.5:280-281)

A specific pitch contour, with *really* in a separate tone unit, might provide greater emphasis, as in (55):

(55) it's ‖∨AL'MOST■ - ‖THR↘EE 'weeks■ ‖R∨EALLY■ (S.2.7:860-862)

One position which occurred only in speech was *really* in clause-final position followed by a comment clause. Such clauses were characteristically uttered within the same tone unit as *really*:

(56) which is ‖QU↘ITE re'vealing R↗EALLY I 'think■ (S.2.9:333)

This contradicts the statement in Quirk et al (1985:1112) that comment clauses, ie disjuncts which express the speaker's attitude to the main clause or his manner of asserting it, generally occur in a separate tone unit.

The clause-constituent separated from the preceding part of the clause by *really* in post-position may be obligatory as in (57) or optional as in (58):

(57) So I'd only need, *r*eally, to begin: "Isn't... (K10 175)

(58) *it ‖doesn't* ▻BOTHER 'me REALLY■ at +‖ALL■ + (S.1.12:506-507)

Post-position was particularly common in cleft sentences and extraposition, where the two parts separated by *really* were connected by copular BE, eg:

(59) it ‖is a PROBLEM ▻really■ ‖how to ΔTEACH this STUFF■ (S.1.9:1128-1129)

It was also frequently found between a head and its post-modifier:

(60) (...) I've ‖no in'tention* NOW of■ - ‖now of • of ▵publishing ▵ARTICLES REALLY■ ‖from • ‖from the BOOK■ (S.3.6:434-437)

The dominating intonation contours of *really* in clause final and post-position were rising and falling-rising. Placed in post-position, *really* typically constituted the second part of a pitch sequence consisting of a fall on the immediately preceding clause element and a rise or a fall-rise on *really*. Except in the cases where *really* was followed by a comment clause (eg example 56), it always occurred at the end of a tone unit (47) or in a separate tone unit (52). In case a comment clause constituted the last part of the utterance, this clause did not carry a nucleus but continued the main pitch contour.

It may be tempting to regard *really* separated from the rest of the clause by a tone unit boundary and/or a pause as an evaluater, reflecting the speaker's attitude to what he is saying, and to regard *really* placed before the head, but with no separating prosodic features, as an intensifier. It is also tempting to take mobility as a criterion of evaluater function. However, this is being much too categorical. Not only must the combined effect of position and prosody be considered but also the wider context (cf Greenbaum 1969:127, 183).

*Really* in clause-final and post-position varied from prosodically unmarked to marked, not only in terms of tone but also other prosodic features, as shown in (61a-d):

(61a) ‖I've been 'working 'pretty ▵HARD 'really■ - (S.2.7:856)

(61b) (...) it's ‖not so 'easy as • you ▵THINK *REALLY■ * (S.1.5:1180)

(61c) (...) it's ‖really for 'you to DECIDE■ ‖REALLY■ --- (S.2.2:24-25)

(61d) I ‖don't know if 'anything's TERRIBLY 'new at ΔALL■ • ‖REALLY■ • (S.2.7:436-437)

In none of these examples does it seem natural to move *really* to initial position as an indication of intensifier function. *Really* here reflects the speaker's view on the entire proposition, only with varying intensity from prosodically unmarked in (a) to a separate tone unit in (d). In (a) the effect resembles that of a communicative cue, eg *you know*; in (d) the effect is that of an after-thought, paraphrasable by 'as a matter of fact'.

The main difference between *really* in clause-final position, as in (61), and in post-position, as in (62), seems to be that it acts more like an evaluater in the first case, with the entire preceding proposition as its scope, and more like an intensifier in the second case, where the nuclear element immediately preceding *really* is placed in focus and gets special emphasis.

(62a) and it's IM||PORTANT 'really■ for ||somebody that has "△more experience than ONE "△SELF■ • (S.2.9:1233-1234)

(62b) but it's ||very im△portant REALLY■ • in ||many {WAYS} to 'write down △what the 'patient COM△PLAINS of■ • (S.2.9:1216-1217)

(62c) (...) I'm ||just not "△INTERESTED e'nough■ • ||REALLY■ *• to ||do THAT■ * (S.3.1:134-136)

(62d) I mean it's ||BEEN 'bad ENOUGH for me as it △IS I 'think■ ||REALLY■ • in ||LOTS of WAYS■ (S.1.9:788-790)

It is still an open question why *really* occurs finally at all instead of inside the clause. The most likely answer is that this is a reflection of the speaker's ongoing planning strategy. He may realize, at that very point, that he wants to modify what he just said, either by softening an assertion, as in (62a), or by giving more emphasis, as in (62c), or by adding an after-thought, as in (62d). There is also the possibility that *really* in post-position can become entirely void of meaning and reflects mannerism rather than a meaningful speaker strategy.

Three functions of *really* were not met with in the written corpus: as re-opener, go-on, and planner. They are all linked to the interactive situation. Re-opener and go-on moves constitute the hearer's response to the current speaker's talk, but with opposite effects on the turntaking system: the re-opener causes speaker-shift while the go-on encourages the speaker to hold his turn. Planners are overt manifestations of the speaker's planning procedure and have a turnholding effect. Re-openers are moves in the interaction which re-open an otherwise completed exchange by eliciting confirmation (cf Stubbs 1983:110 and Stenström 1984a:240-241). They are often realized by comment-Qs, such as *did he* or *aren't you*, or by phrases involving *really*, eg *did he really go there*. But *really* on its own, indicating surprise, incredulity, etc, may have the same response-eliciting effect. A re-opener is either the hearer's reaction to an inform (Stenström 1984a:86) as in (63), or the questioner's reaction to the response as in (64):

(63) B: (...) and ||THÀT was■ - you ||KNÓW■ in ||times that ΔÌ can re'member■ •
A: *||RÉALLY■ *
a: *good* Lord
B: ||oh YÊS■
A: ||YÈS■ (S.2.3:34-42)

(64) A: ||Oscar is ΔGǑING to the ▻States■
B: «||WÈLL■ » this is what I "||HĚARD■ just be||fore I came AΔWǍY■ ---
A: ||RÊALLY■
B: «||YÈS■ » -- (S.1.2:349-354)

Informs, as in (63), are generally followed by a feedback item (cf Coulthard 1981:25), ie a minimal answer like *mhm* or *I see*, indicating that the hearer has received the information, but in this case the hearer queries the truth of the message, and *really* serves as a request for confirmation.

The normal pitch contour for *really* as a re-opener is falling-rising or rising. The fact that *really* in (64) with a rising-falling tone still functions as a re-opener is probably an effect of the long pause after the response. A contributory factor may be that *really* retains some of its original meaning and automatically serves to check the truth/falsehood of the preceding utterance when occurring in this particular position.

*Really* as a go-on is less expressive than *really* as a re-opener and does not invite a confirming response. The current speaker is encouraged to continue, and there is no speaker-shift. Compare (65) with (64):

(65) B: I ||don't know if he ▻DRÒPPED 'that■ --
A: ||oh RÊALLY■
B: cos • well ||I ||I don't ΔKNǑW■ *||when he was 'trying to FÌND* (S.1.5:257-260)

It is highly probable that the present speaker registers the listener's feedback but, contrary to what was the case in (64), he does not show it by a confirming response.

The term 'go-on' is equivalent to 'continuer' used by Schegloff (1982) for items like *uh huh*, by which the hearer passes up the opportunity of taking over the turn. By inserting the continuer *uh huh* the hearer shows that he understands and is paying attention to what is being said and, above all, that he is aware of the current speaker's intention to keep on talking. However, Schegloff does not include *really*, which he refers to as a sort of 'reaction' invited by the immediately preceding talk, 'aside from, instead of, or in addition to the continuer'. But the present data shows that *really* often has the same effect as *uh huh*.

*Really* as a go-on typically carries a falling or falling-rising tone. Whether *really* should be interpreted as a go-on or as a re-opener depends on the way B's utterance is understood:

(66) B: (...) and pre||sumably he's 'got ▻something ▵equally ΔFÀTAL■ ||or perΔhaps it ΔÌS 'lung 'cancer■ •

A: ||RÉALLY■

B: ||this is 'all 'very ▵SÀD■ I ||feel • BÀD about «that»■ (S.1.4:1042-1046)

*This is all very sad* can either serve as a confirmation invited by *really* or as a comment that speaker A would have added anyway. In the first case, *really* would be a re-opener, in the second a go-on.

*Really* as a planner is used as a strategic device in the planning of speech, sometimes characterized as an empty 'filler' (cf Brown 1977:107 ff) and equivalent to a pause, sometimes as a 'projecter', ie a temporary substitute for a not yet specified intensifier or an intensifier waiting for a head, and sometimes as an emotionally expressive 'react' signal, equivalent to an 'evaluater'. In each of these functions, *really* tends to occur in a hesitation area, either at the beginning of an utterance where the speaker has not yet made up his mind how to continue, or within the utterance when he suddenly loses the thread or otherwise stops to reconsider (see also Chapter 8).

The functions are manifested somewhat differently: *really* as a filler does not generally carry a nucleus, but as a react it is always prosodically marked with some prosodic feature reflecting its degree of intensity. In its filler function *really* is neither oriented towards the preceding speaker's utterance nor towards an element that follows in the current speaker's utterance; as a react it is either oriented towards the preceding speaker's utterance or towards an element in his own utterance; as a projecter it points forward, looking for something to modify.

That *really* can be used as a planner is obvious both in the filler and in the projecter functions. *Really* in (67) is pointing forward to a head later on in the utterance:

(67) and ||I • ||I ▻get ∆really ▻[əm] -- «you ||know» when ∆[?]when I'm 'trying to CÒOK■ • and ||people come and CHÀT■ I ||I get ∆terribly put ÒFF■ - (S.2.7:69-71)

It is therefore different from the filler in (68) where *really* can hardly be described in syntactic terms at all but is outside the syntactic structure of the utterance (cf Brown 1977:109):

(68) but [ə] ||really I've got about • ▵THRÈE WÉEKS■ "||less than THÁT■ of ||hardish ▵WORK■ (S.1.1:155-157)

My suggestion is that *really* in (67) is part of the planning process in that it anticipates a head, here realized by *put off*; that this is so seems to be confirmed by the fact that it is replaced by the intensifier *terribly*. In anticipating a head *really* acts as a 'dangling' intensifier. At the same time it attracts the hearer's attention to the fact that the real message is still to come and also acts as a turnholder. *Really* as a filler in (68) fits Brown's description of fillers, which says that their principal duty is 'to fill the silence and maintain the speaker's right to speak, while he organizes what he wants to say' (1977:109).

*Really* frequently cooccurs with various kinds of hesitation phenomena: unfilled and filled pauses, softeners, repetitions, reformulations, new starts, hedges (*kind of*, *sort of*) and other fillers (*well*) (see (69) and further Svartvik & Stenström 1985):

(69) and ||ÀLL this was DÓNE [ə:]■ -- ||by --▻kind of ▻letting - [ə:] -- • ||{WÈLL} RÈALLY by 'just [ə:] -- 'sort of [ə]■ - ||starting from ∆NÒTHING■ (S.2.3:115-117)

The speaker probably knows right from the start what he wants to say but not how to put it in words. In (70) *really* is met with in its react function. It occurs in a separate tone unit and is similar to an evaluater at the syntactic level:

(70) d: you're you're an awkward customer aren't you

A: (--- laughs) - [ə:m] --- ||{WE̖LL} 'OAK̂■ ||RĚALLY■ - I ||mean ---

(S.2.4:813-816)

Why, then, is *really* so much more common in speech than in writing? One reason is its versatility: not only can it be used to emphasize different parts of an utterance as well as the entire utterance, but it can also be used for various interactional purposes. Another reason is that the functions of *really* may be neutralized to the extent that it can be used to fill empty gaps in speech, sometimes for the sake of stalling, sometimes even for rhythmical reasons. There are other adverbs that can serve some, but not all, of these functions. *Actually* is perhaps the nearest candidate but it cannot be used as an intensifier, nor can it serve any of the interactional functions.

## 5.6 The *right* set

The D-items *right*, *all right* (sometimes spelled *alright* in the transcription), and *OK* (or *okay*) have the following general characteristics (cf Schiffrin's conditions for 'discourse markers' 1986:328):

They do not serve as elements of clause structure.
They serve several functions in the discourse.
They operate at more than one discourse level.
They occur frequently at the beginning and end of turns.
They generally appear in a separate tone unit.

When examining the use of these items (which for convenience will henceforth be referred to as the '*right* set') in LLC, I looked for such features as text frequency, position and function in the discourse. When comparing their frequency in relation to other discourse items I restricted myself to items with a potential response function, including feedback (see Table 5:2). Since most of them can also serve functions analysable in syntactic terms, I have indicated D-function ratio, ie discourse function in relation to grammatical function (expressed in per cent).

*Yes/yeah* outnumbered the rest, and the high frequency of positive signals is not difficult to explain. First of all, a smooth conversation requires cooperative partners, which means that *no* was a rare response compared to *yes*. Second, conversation being a mutual undertaking, the listening party is aware

**Table 5:2**. Typical D-items in the London-Lund Corpus.

| ITEM | D-ITEM RATIO (%) | D-ITEM FUNCTION Frequency | TOTAL ITEM Frequency |
|---|---|---|---|
| *yes/yeah* | 100 | 4263 | 4263 |
| *mhm* | 100 | 1621 | 1621 |
| *oh* | 100 | 1519 | 1519 |
| *that's right* | 100 | 79 | 79 |
| *that's OK* | 100 | 13 | 13 |
| *that's all right* | 100 | 9 | 9 |
| *OK* | 92 | 249 | 270 |
| *well* | 86 | 2675 | 3103 |
| *right* | 55 | 411 | 740 |
| *all right* | 48 | 116 | 242 |
| *indeed* | 30 | 56 | 184 |
| *really* | 15 | 115 | 780 |
| *quite* | 13 | 106 | 830 |
| *of course* | 11 | 66 | 616 |
| *certainly* | 5 | 13 | 211 |
| *probably* | 2 | 7 | 316 |

that he is required to show (or at least pretend) that he is an attentive and interested listener by inserting 'backchannel items' at more or less regular intervals. Such backchannel items were frequent and mostly realized by *yes* (*yeah*) or simply *mhm*, the third most frequent item in the list. Other frequent items were *well* and *oh*, which were both used mainly as response initiators, *well* typically introducing an 'insufficient' response (Lakoff 1973, Stenström 1984b) and *oh* typically used as an 'information receipt' (Heritage 1984) or as a 'reinforcer' (Stenström 1984a:147). *Well* as a 'staller' (p 141) is discussed in Svartvik (1980a), Carlson (1984), Schourup (1985), and Schiffrin (1986). Functions of *oh* have been described by, among others, James (1972) and Aijmer (1987).

Also *quite* and *really* were found more often than the *right* set (but notice the difference in D-function percentage). *Quite* and *really* as responses are illustrated in Quirk et al (1985:612, 628). (For *OK*, see Merritt 1984:139-47.) *Indeed*, with a lower D-frequency than *quite*, *really*, and *of course*, was still used more often as a response signal (30 %). Like *quite* and *of course*, *indeed* served as an 'emphasizer' in most of the cases, adding to the force of the response proper (as in *yes indeed*). (Note that 'emphasizer' is used as a discourse term and not as a grammatical term as in Quirk et al 1985:485.) In the few cases where *probably* and *certainly* served as response signals they were also generally used as emphasizers. The typical realization of emphasizers was *that's right*, which was added to the list of response (feedback) items for comparison, like *that's all right*, *that's OK*, and *it's all right*. Due to their restricted use, as smooth-overs after apologies, the latter three were extremely rare.

What is of particular interest, when it comes to the *right* set, is that each one of the D-items served a wider range of discourse functions than any of the other items mentioned, despite a much lower frequency than some of them (cf Stenström 1987:94). Another noticeable feature is that they were all used very often in telephone conversations, especially *OK*, which turned out to be the typical telephone device (see Table 5:3). Only *oh*, *yes/yeah*, and *mhm* were used exclusively as interactional devices. One example is (71) where *oh* is used as a follow-up move in a questioning exchange by which speaker D indicates that he has received the information:

(71) D: and ||what did you 'do your •

A: ||SÒRRY ■

>D: DE△GRÈE • in ■

A: in ||music and ÈNGLISH ■

D: ||ÒH ■ (S.1.5:1293-1296)

The other items in the list were used also in various grammatical functions; *right* as a noun:

(72) exercised his right as Chief of STÀFF

**Table 5:3.** The *right* set in face-to-face and telephone conversation.

| ITEM | FACE-TO-FACE | % | TELEPHONE | % | TOTAL |
|---|---|---|---|---|---|
| *right* | 93 | 34 | 177 | 66 | 270 |
| *OK* | 27 | 11 | 222 | 89 | 249 |
| *all right* | 53 | 46 | 63 | 54 | 116 |
| TOTAL | 173 | 27 | 462 | 73 | 635 |

as an adjective:

(73) on my right HÀND

as an adverb:

(74) right up to the ÀRCH

*all right* and *OK* as adjectives:

(75) you know it's all RǏGHT

(76) it may not be ÓK to ÈVERYBODY

and as adverbs, ie as emphasizer subjunct (cf Quirk et al 1985:587):

(77) she can go shopping all right

(78) make sure that he's THÈRE ÓK

and process adjunct:

(79) I hope that drive goes ÒK

In Section 5.2, I demonstrated how the function of turns varies with their position in the exchange, and I pointed out that the function of discourse items also varies with their position in the turn. The different positions in the turn will be referred to as follows:

**Slot 1**: First item in the turn

(80) ||ALRÍGHT■ ||LÒOK■ • ||[ə:]ΔI'm Robin BÚSS■ (S.8.2:894-896)

**Slot 2**: Second item, following another D-item in slot 1

(81) ||ÒK■ *||RÌGHT■ * (S.2.11:1369-1370)

**Slot 4**: Last item in the turn

(82) and ||I'll TÀLK to YOU■ – ||like at that 'time for a'bout FÌNANCE ÁLRIGHT■

(S.8.2:839-840)

**Slot 3**: An item between slots 1 or 2 and 4

(83) ||nearer the two THÌRTY■ ||ALRÌGHT■ ||well we (S.8.2:1228-1230)

**Table 5:4**. Distribution of the *right* set in the turn.

| ITEM | SLOT 1 | SLOT 2 | SLOT 3 | SLOT 4 | SEPARATE TURN | TOTAL |
|---|---|---|---|---|---|---|
| *right* | 114 | 42 | 24 | 20 | 70 | 270 |
| *OK* | 77 | 50 | 18 | 40 | 64 | 247 |
| *all right* | 39 | 18 | 18 | 24 | 17 | 116 |

The distribution of the *right* set in the turn is shown in Table 5:4. It is notable that both *right* and *OK* occurred more than twice as often as *all right*. Moreover, all three items were more common within the turn than in a separate turn, and they were most often found at the very beginning of the turn. The figures also indicate that we can expect *right* more often than *all right* and *OK* in a separate turn and in slots 1 and 3, and *OK* more often than *right* and *all right* in slots 2 and 4.

Position in the turn was found to correlate with specific discourse functions (see Figure 5:2; for exact figures see Stenström 1987:95). We notice that go-on moves, re-openers, and follow-ups generally appeared in a separate turn, while uptakes, responses, call-offs and closes were turn-initial, and emphasizers were found in the second slot. Frames usually appeared in turn-medial position, whereas prompters and questions came at the end of the turn. As for the realization of these functions, see Table 5:5.

We find that *right* was more often used as a follow-up and a close than *all right* and *OK*; that *all right*, despite its low frequency, was the typical prompter and re-opener; and that *OK* served as a call-off, as a question, and as a frame more often than *right* and *all right*. Examples (84) - (91) illustrate what was typically done in each of the four positions and which item typically occurred in that position.

**Figure 5:2**. Typical functions in relation to position in the turn.

| SLOT 1 | SLOT 2 | SLOT 3 | SLOT 4 | SEPARATE TURN |
|---|---|---|---|---|
| Uptake | Emphasizer | Frame | Prompter | Follow-up |
| Response | | | Question | Go-on |
| Call-off | | | | Re-open |
| Close | | | | |

**Table 5:5.** Item and function.

| FUNCTION | *RIGHT* | *OK* | *ALL RIGHT* | TOTAL |
|---|---|---|---|---|
| Response | 45 | 68 | 22 | 135 |
| Frame | 27 | 36 | 26 | 89 |
| Follow-up | 53 | 28 | 3 | 84 |
| Emphasizer | 38 | 31 | 13 | 82 |
| Call-off | 19 | 55 | 7 | 81 |
| Close | 33 | 12 | 10 | 55 |
| Prompter | 9 | 2 | 16 | 27 |
| Uptake | 23 | 2 | 8 | 33 |
| Go-on | 19 | 1 | 4 | 24 |
| Question | 1 | 11 | 2 | 14 |
| Re-open | 3 | 3 | 5 | 11 |
| TOTAL | 270 | 249 | 116 | 635 |

The separate turn proved to be the characteristic position of various types of feedback in a broad sense, namely follow-ups, go-ons, and re-openers. It contained a follow-up move more often than any other move:

(84) A: ||shall we 'keep those ▵brackets as they ÀRE■ –

B: ||YÈS■

A: ||RÌGHT■ (S.9.1:515-517)

The follow-up is generally the questioner's evaluation of the response in a questioning exchange as in (84), but may sometimes serve as the addressee's evaluation of the inform in a non-questioning exchange. It terminates the exchange unless the termination is temporarily postponed by a re-opener.

The go-on move is an even more typical feedback device:

(85) C: ||and the ΔÓTHER■ • ||[ði: əm] • her ΔSÀLARY■ ||{WÀSN'T} ▵QUÌTE ENÓUGH■ to ||meet ['ði 'ði ði] re▵quirements of the "▵BÙILDING SOCÍETY■ ||so *Δshe*

B: *||NÒ■ *

C: fell 'down on ▵FÌNANCE■ •

B: *‖RÌGHT■ *

C: *so I've* ▻got just △two LÈFT■ (S.8.1:1169-1176)

The go-on is the listener's minimal feedback by which he assures the current speaker that he is listening. Since this move is inserted while the other party is talking, it often causes a temporary break in the syntactic continuity of his talk, but always without causing a speaker-shift.

*That's right,* sometimes alternating with *right* for emphatic reasons, was characteristically used as a go-on when the current speaker's statement referred to an event known to both A and B, while *all right* was used when the utterance referred to an event known only to B. The main difference between the follow-up and the go-on is that the latter does not involve speaker-shift while the former terminates the exchange.

The re-opener reflects a higher degree of involvement than the go-on move does:

(86) B: ‖oh I △SÈE■ - oh well 'we can [ə:] - we can ‖give 'that 'to her a'mong △ÒTHER things 'then■

A: ‖YÈS■ •

B: ÁLL right■

A: ‖YÈS■ and I'll ‖come via HÀRRODS to YÓU you see■ and ‖see if I can 'get those △SÒCKS {for ‖ÌAN■ }■

B: ‖OK■

A: ‖R/ÍGHT■

B: ‖RÌGHT■ and ‖keep an 'eye out for 'something 'for ▻for • ÈMMELINE■ (S.7.2:137-147)

The re-opener is used to query unexpected information or ask for confirmation. In (86) *all right* re-opens the first exchange which would otherwise be terminated by *yes*. *Right* re-opens the second exchange which would otherwise have been terminated by *OK*. The tone of the follow-up and go-on moves was generally falling, that of the re-opener almost always rising (cf *really*, p 156).

D-items occurring in turn-initial position (slot 1) generally either responded to the previous move or served as a link between the previous and the immediately succeeding move. This is where we find first response acts and uptakes. The term 'response' denotes answers to questions as well as retorts to non-questions, both of which often consist of a first response act followed by a second, optional act:

(87) B: I'll ||get her to 'ring you 'when she comes ÌN■

A: ||ÓK■ ||RÌGHT■ (S.9.1:220-221)

The response move consists of *OK* which accepts the offer expressed by B and the emphasizer *right*. The tone of the response could be either rising or falling. Although the choice is a matter of finality, it seems reasonable to assume that a rising tone reflects a more favourable attitude than a falling tone.

'Uptakes' look back and create a link with the previous speaker's move (cf Edmondson 1981:84):

(88) B: ||YÈS■ – ||I mean if ▵people 'take it SÈRIOUSLY■ ·

c: *[m]*

A: *||RÌGHT■ * well "||I've been 'shown up to be a com▵plete PHÌLISTINE■

(S.2.10:579-583)

Speaker A uses *right* to validate what speaker B just said before continuing. Uptakes and follow-up moves in turn-initial position provide a very similar type of feedback. The main difference between them is that the speaker who produced the uptake goes on and initiates a new exchange in the same turn; the speaker who produced the follow-up usually terminates the exchange while the next speaker initiates the new exchange.

Call-offs and closes, which were also characteristically found in slot 1, will be dealt with below in connection with telephone conversation.

Slot 2 was typically occupied by the 'emphasizer', as illustrated in (87) where *right* in slot 2 emphasizes *OK* in slot 1. In this example the emphasizer accompanies a response act. In other cases it emphasizes a follow-up or a go-on. Although emphasizers are optional they were more often than not added to first acts realized by the *right* set in response moves (see also Stenström 1984a:233 for *yes/no* responses). Emphasizers usually occupied a separate tone unit and carried a rising more often than a falling tone (especially when realized by *that's right*).

Slot 3 was the position of frames:

(89) ||THAT'S it■ the ||FOLKLORE {SO||CIETY 'library■ }■ ||YES■ ||that's ||that's RIGHT■ ||that's «FINE»■ • ||YEAH■ – "||RIGHT■ • [ə:m] • ||well NOW■ [jə] you ||you SAY in OTHER words■ (S.3.3:217-225)

Frames mark a boundary in the discourse and signal the transition between two stages (Sinclair & Coulthard 1975:44, Stenström 1984a:125). In cases where the frame occurred turn-initially, it either marked the beginning of a new transaction (as in mid-position) or introduced the very first thing a speaker said, or marked the return to, and reconfirmation of, an arrangement agreed upon earlier (*all right*), or marked the resumption of an interrupted narrative. *Right*, generally with a falling tone, proved to be the characteristic 'switch-off' signal (signalling end of topic), whereas *all right* (usually with a rising tone) was the characteristic 'switch-on' signal (signalling new topic).

Slot 4, finally, was the typical position of the prompter:

(90) B: ||try to ▷READ it■ as ||if you're ▷not • *YOURSELF■ • ||that's*
A: *as ||if it △wasn't MINE■ • ||all RIGHT■ *
>B: RIGHT■ • ||YES■ -- ||read from 'there to the **END■ ** (S.3.1:265-271)

The prompter transforms the statement to which it is attached into a request for confirmation or acknowledgement. In this function, the D-item generally constituted a separate tone unit, usually followed by a pause. Occasionally, *right* and *all right* in slot 4 acted rather like questions. One example is (91):

(91) ||CLIVE■ ||RIGHT■ --- (S.9.2:621-622)

This example, which occurred at the beginning of a telephone call, can be paraphrased as 'Is that Clive speaking?'. In such cases the tone was always rising.

Notice that *all right*, the overall least frequent item, occurred as often as *OK* in turn-medial position (slot 3) and more often than *right* in turn-final position (slot 4). The most common functions of *all right* were frame, response, and prompter, in that order.

The different positional distributions in face-to-face and telephone conversation can be seen in Table 5:6. The difference is particularly obvious in the separate turn and in slot 1, where the high figures in telephone calls result from the use of, especially, *right* and *OK* in closing sections (Schegloff &

**Table 5:6.** Distribution of the *right* set in face-to-face and telephone conversation.

| | SLOT 1 | | SLOT 2 | | SLOT 3 | | SLOT 4 | | SEPARATE TURN | |
|---|---|---|---|---|---|---|---|---|---|---|
| | face | tele | face | tele | face | tele | face | tele | face | tele |
| *right* | 38 | 76 | 14 | 28 | 16 | 8 | 10 | 10 | 14 | 56 |
| *OK* | 3 | 71 | 5 | 45 | 12 | 6 | 0 | 41 | 3 | 61 |
| *all right* | 12 | 27 | 8 | 10 | 12 | 6 | 9 | 15 | 5 | 12 |

Sacks 1973) where *OK* was the typical first pair part (call-off) and *right* the typical second pair part (close).

There are two types of closing exchange: the potential 'pre-closing', which occurs at the end of a phase in the dialogue and provides an opportunity for a new topic to be introduced instead of the closing, and the 'closing proper' which terminates the conversation (Schegloff & Sacks 1973). In this data the two types of closing were usually merged into one:

(92) A: «we'll» ||want at ▻least 'half an ▵HOUR■

B: ||UHUH■

............

A: O||K■ Call-off

B: ||*RIGHT*■ ||{OH} GREAT■ Close

A: ||OK THEN■

B: ||see you THEN■

A: ||GREAT■

B: bye ||BYE■ (S.7.3:227-236)

*OK*, the first closing move, is repeated by *OK then*, before *bye* terminates the closing section.

The *right* set can occur in any of the turns in the exchange and in any of the slots in the turn. Does this also mean that one item can substitute for another in the set without a change of meaning and function? Could, for example, any one of the three D-items *right, all right*, *OK* occur in the following exchanges?

**Request / accept:**

(93) could you hold on

**Thanks / response:**
(94) thanks very much

**Suggest / agree:**
(95) shall we turn to the applicants now

**Direct / accept:**
(96) make sure there's not a question about that

**Inform / follow-up:**
(97) it's under H for Harry

*All right* and *OK* were used more frequently than *right* after requests, suggests, and directs, whereas *right* and *OK*, but not *all right*, were used after *thanks* in the data. *Right* was the typical follow-up in informing exchanges. All three items were used as a follow-up move in questioning exchanges:

(98) A: ||shall we 'keep those ▵brackets as they ARE■
B: "||Y\ES■
A: ||RIGHT■ (S.9.1:515-517)

(99) A: [ə:] ||shall I come EARLIER■ or at ||four o'CLOCK■
B: ||no I should ▵COME {at ||four o'*CLOCK■ }■ «if you»*
A: *||all RIGHT■ * (S.7.2:167-169)

(100) A: and I'll ||post ▵OFF 'Rita's 'parcel SHALL I■ •
B: [ə] ||YES■ ||COULD you■
A: ||OK■ • (S.7.2:100-103)

Are *right*, *all right*, and *OK* interchangeable here, or does the type of questioning exchange determine which item realizes the follow-up move? The first question can be paraphrased by 'do you think we should?', the second and third questions by 'would you like me to?'. *Right* in (98) is equivalent to 'I see', *all right* in (99) and *OK* in (100) are equivalent to 'I accept'. The question of interchangeability is extremely complex. Not least does it involve the particular way things are said that cannot be marked graphically in the transcript. Therefore, I am not ready to draw any conclusions on this point.

Summing up, the members of the *right* set were not the most frequent discourse items in the corpus but they were found to serve the widest range of functions. All three occurred more frequently in the telephone calls than in the face-to-face conversations. *OK*, the typical telephone device, was particularly common as the call-off move in the closing section of the call, while *right* realized the close move.

When identifying the functions of the set in the dialogues, I took their position in the exchange structure as a starting-point. All three were found to occupy the same positions in the exchange, to fill the same slots in the turn, and to serve the same functions, but with different frequency.

Items providing feedback of various kinds were found to make up entire speaker-turns. Items serving as a response to what was said in the previous utterance or linking two successive utterances were found in utterance-initial position. Next came emphasizers which gave more force to utterance-initial items. Items marking topic boundaries and signalling transitions were found in the middle of an utterance, and items appealing for feedback or eliciting a response at the end.

*Right* was the typical realization of the follow-up move, *all right* that of the prompter, and *OK* that of the response. Although all three items were found to occupy the same positions, they were not fully interchangeable. The choice of D-item was related to type of mode (face-to-face or telephone), type of exchange (eg informing or requesting), and type of preceding move.

## 5.7 Towards a model of analysis

A model of analysis may not be indispensable for the interpretation and description of discourse items, but it is certainly helpful. Exactly what aspects the model should be able to handle and how detailed it should be is a different matter. I have aimed at a model (described in Stenström 1989) that can handle not only the various levels of a conversational exchange but also bridge the gap between grammar and discourse by indicating when items that are traditionally referred to in grammatical terms have a predominantly interactive function.

It is possible to identify three different types of D-items:

[1] PURELY INTERACTIONAL, including lexical items which cannot be described in terms of clause elements;

[2] MAINLY INTERACTIONAL, including lexical items that are primarily used as interactional devices but may be used, in some environments, as clause elements;

[3] ALSO INTERACTIONAL, including adverbials of various classes used as interactional or discourse organizing devices.

The three categories can be seen along a 'clause-integration' scale, which implies that the more interactional an item is, the less integrated it is - and vice versa. The most common D-items in the data are listed in Figure 5:3.

All the items, except those in the first group, are provided with at least two word-class tags which indicate their potential interactive and/or grammatical function. The word-class labels differ from those used earlier in this chapter but follow the TESS tagging dictionary (see pp 101ff), which was based on Quirk et al (1985) and introduced for the automatic word-class tagging. However, the three groups suggested here do not agree entirely with their categorization. They regard *ah*, *aha*, *mhm*, and *oh* (group 1) together with eg *ouch* and *pooh* as 'purely emotive' interjections which 'do not enter into syntactic relations'; *yes* and *yeah* (group 1) and *all right* (group 2) are referred to as formulae, most of which, are 'used for stereotyped communication situations' and can be analysed into clause elements in a very limited way (1985:852-53). *Ah*, *mhm*, and *oh* do not enter into syntactic relations, but I do not see them as purely emotive, and I find it difficult to look upon *yes*, *yeah*, and *all right* as formulae. Nor do I think that any of them are used specifically in stereotyped communication situations.

The word-class membership of D-items (including single words as well as strings of words used as D-items) is shown in Figure 5:3. Although they are all classed as adverbials (indicated by A in the labels), the 'purely interactional' items (group [1]) do not really belong to any traditional word-class category, unless we regard all items that do not fit the definition of other word-classes as adverbs (cf Quirk et al 1985:438). By contrast, the items referred to as 'mainly interactional' (group [2]) and 'also interactional' (group [3]) are provided with at least two word-class labels indicating their potential functions at the discourse and/or grammatical level.

Purely interactional items do not normally serve as clause constituents. Items that are not purely interactional are more often found in a context where they serve as elements of clause structure. D-items realized by sentence fragments and simple sentences are the only ones that can always be analyzed as clause elements.

The question is now: what kind of discourse model can handle not only the interactional aspect manifested in the exchange but also the grammatical aspect manifested in the sentence? In Stenström (1989) I discuss a combined, two-level model which can handle both the interactional function of D-items in a speaker's turn and their potential syntactic functions (as indicated in Figure 5:3) as well as their syntactic structure in appropriate cases (eg *that's right*, *I*

**Figure 5:3**. Categories of interactional D-items (see pp 101ff for labels).

| [1] PURELY INTERACTIONAL | |
|---|---|
| *ah* | AQres |
| *aha* | AQres |
| *mhm* | AQres |
| *oh* | AQres |
| *yeah* | AQpos |
| *yes* | AQpos |

| [2] MAINLY INTERACTIONAL | | | | | | |
|---|---|---|---|---|---|---|
| *no* | AQneg | B3neg | | | | |
| *please* | AQpol | VA | | | | |
| *I see* | AQres | BHsub+VA | | | | |
| *I mean* | AQsof | BMsub+VA | | | | |
| *you know* | AQsof | BHper+VA | | | | |
| *you see* | AQsof | BHper+VA | | | | |
| *OK* | AQres | AQfra | | | | |
| *all right* | AQres | AQfra | JA | | | |
| *thank you* | AQtha | VA+BHper | | | | |
| *that's right* | AQres | RC+VB3+JA | | | | |
| *that's all right* | AQres | RC+VB3+JA | | | | |
| tag Qs | AQtag op(+not)+JA | | | | | |
| *well* | AQwel | ASint | JA | NC | AC | |
| *sure* | AQres | ADcnt | ASint | JA | ASemp | |
| *right* | AQres | AQfra | AApro/spa | JA | NC | VA |

| [3] ALSO INTERACTIONAL | | | | |
|---|---|---|---|---|
| *anyway* | ACcon | AQfra | | |
| *in fact* | ACcon | AQfra | | |
| *maybe* | ADcnt | AQres | | |
| *perhaps* | ADcnt | AQres | | |
| *probably* | ADcnt | AQres | | |
| *absolutely* | ASint | AQres | | |
| *however* | ACcon | AQfra | | |
| *of course* | ACcon | ASemp | AQres | |
| *certainly* | ADcnt | ASemp | AQres | |
| *obviously* | ADcnt | ASint | AQres | |
| *indeed* | ADcnt | ASint | ADcon | AQres |
| *really* | ADcnt | ASint | ASemp | AQres |
| *honestly* | ADsty | AApro | ASemp | AQres |
| *now* | AAtim | ACres | CCtim | AQfra |

*see what you mean*). As I try to demonstrate, a model of this kind has many advantages. The most obvious disadvantage is that it tends to become rather bulky, especially if we want to make an extensive analysis, not only of individual D-items but of entire turns.

## Note

Section 5.5 on *really* is based on Stenström (1986) and Section 5.6 on the *right* set is based on Stenström (1987).

6

# Spoken English and the dictionary

*Bengt Altenberg*

## 6.1 Introduction

On the whole, dictionaries tend to reflect the written language rather than the spoken.[1] We only have to take a cursory glance in a dictionary to notice this. The written idiom dominates, and many words and expressions that are characteristic of speech are either missing or inadequately described. Even recent and ambitious dictionaries which put special emphasis on 'real' language as it is used in 'natural communication', reveal a lingering written bias that is sometimes irritating to a user interested in the spoken word.

Perhaps this written bias is only what we can expect: speech has only recently come of age as a variety worthy of scholarly attention, and our knowledge of the spoken language still lags behind that of the written. Perhaps it also reflects the difficulty of describing speech in traditional terms - especially the 'ungrammatical' variety we hear in everyday conversation, with its unplanned 'messiness', its interactive and emotive character, and its reliance on intonation and gesture to convey meanings.

Yet, our knowledge of natural spoken discourse has made substantial advances in the last decades, and we are now beginning to see more clearly the inadequacies of our dictionaries in their treatment of the spoken medium. I will here touch on two areas where I think contemporary dictionaries fail to give an adequate representation of speech: the use of intonation to differentiate adverbial functions and the treatment of certain speech-specific 'discourse items'.

I will use two learners' dictionaries to illustrate my points: the new edition of the *Longman dictionary of contemporary English* (*LDOCE*) and the *Collins COBUILD English language dictionary* (*COBUILD*), both published in 1987. The choice of these was natural. They provide the most up-to-date and authoritative accounts of contemporary English vocabulary and are probably the most influential and useful monolingual dictionaries currently available to learners and teachers of English. Moreover, both focus on 'ordinary everyday English' drawn from extensive corpora of citations or text samples, a fact which should guarantee an exhaustive and reliable treatment of the spoken word.[2]

Although both *LDOCE* and *COBUILD* are learners' dictionaries, I will not examine them from the user's point of view or evaluate them as tools for language learning. Rather, my chief concern has been a more general linguistic phenomenon: the difficulty of describing speech in terms primarily developed for writing. This difficulty is familiar to anyone who has tried to make a grammatical analysis of genuine conversational data. My reflections here spring from this kind of experience, and in particular from the descriptive problems we have encountered within the TESS project in our efforts to write formalized rules for automatic grammatical analysis and intonation assignment on the basis of the London-Lund Corpus.

## 6.2 Adverbs and intonation

The fact that speakers make use of intonation (as distinct from the pronunciation of individual words) to express their intentions means that they can make functional distinctions that are difficult or impossible to express in writing. Adverbs illustrate this point particularly well. Many adverbs have multiple functions as manner adverbial (adjunct or subjunct in the terminology of Quirk et al 1985) or sentence adverbial (conjunct or disjunct). *Briefly*, for instance, may be used both as a manner adjunct and as a style disjunct (the latter expressing the speaker's comment on the form of his utterance), as illustrated in the following examples:

(1) we discussed the matter BRIEFLY■

(2) BRIEFLY■ there is nothing more I can DO about it■

As a manner adjunct (1), *briefly* normally occurs clause-finally with nuclear prominence, but as a style disjunct (2) it is typically placed clause-initially in a separate tone unit, usually with a falling-rising tone (cf Allerton & Cruttenden 1974:21 and Bing 1984:16ff). *LDOCE* describes *briefly* (under the adjective *brief*) as follows:

> **~ly** *adv*: *The President stopped briefly in London on his way to Geneva. Briefly, I think we should accept their offer.*

This description is not particularly illuminating. A functional distinction is implied (by word order and punctuation), but there is no attempt to make the distinction explicit. The treatment of *frankly*, which has the same functional possibilities as *briefly*, is hardly more adequate; the adverb is given a separate entry and described as follows:

> *adv* 1 in an open and honest manner 2 speaking honestly and plainly: *Frankly, I don't think your chances of getting the job are very good.*

Here, a functional difference is suggested (by the two definitions), but only one function is illustrated. Moreover, the important role played by word order and intonation remains unclear.

*COBUILD*'s treatment of these adverbs is more helpful. Both are presented in separate entries and their different functions are carefully explained in numbered paragraphs and illustrated with several examples. *Briefly*, for example, is described as follows:

> **1** Something that happens or is done **briefly** happens or is done for a very short period of time. EG *He smiled briefly... 'Good morning, Tommy,' he said, looking up briefly.*
>
> **2** If you say something **briefly**, you use very few words or give very few details. EG *She told them briefly what had happened... Robertson answered briefly and without interest... Put briefly, his argument was this.*
>
> **3** You can say **briefly** to indicate that you are about to say something in as few words as possible, rather than giving a full description or explanation or to indicate that you are about to summarize what you have just been talking about. EG *The facts, briefly, are these... For reasons beyond my control (briefly, money) I once moved five times in eighteen months.*

In addition, semantic and grammatical information is supplied in an extra marginal column in the form of synonyms, antonyms and/or superordinate terms and a grammatical code system based on a mixture of structural and positional criteria. Thus, the first two senses of *briefly* are classified as 'ADV AFTER VB' (adverb only used after a verb or a verb + object) and the third as 'ADV SEN' (sentence adverb with various specified positions).

*Frankly* is divided into two senses, the first classified as 'ADV SEN', the second as 'ADV WITH VB' (adverb modifying a verb and occurring either before or after the verb (plus object, if any)):

> 1 You use **frankly** when you are stating an opinion to emphasize that you mean what you are saying, even though the person you are speaking to may not like it. EG *Frankly, this has all come as a bit of a shock... It is frankly absurd... Quite frankly, I am too miserable to care.*
>
> 2 If you say or do something **frankly**, you say or do it in an open, honest, and straightforward way. EG *He asked me to tell him frankly what I wished to do.*

As these examples show, *COBUILD*'s description is more detailed and explicit than *LDOCE*'s. This is partly achieved by greater elaboration of the entries, but also - and more interestingly - by its adverbial coding system. This system, which distinguishes altogether five adverbial functions, is a useful innovation in the lexicographic treatment of adverbs. (Apart from the categories mentioned, two other functions are recognized: 'ADV + ADJ/ADV' = modifiers of adjectives and adverbs, and 'ADV BRD NEG' = 'broad' negatives like *hardly*, *scarcely*, *seldom*, etc.) Yet, despite its usefulness, some functional differences are still broadly suggested rather than sharply defined. The distinction between 'ADV AFTER VB' and 'ADV WITH VB' is blurred by functional and positional overlap (mainly due to the heterogeneity of the latter class), and the category 'ADV SEN' comprises a wide range of sentence adverbs (conjuncts and disjuncts) whose positional characteristics can only be captured in rather general terms: 'usually placed at the beginning of a clause followed by a comma or in the clause separated by commas. A few come at the end of the clause.' Hence, the distinctive position(s) of these adverbs must be inferred from the illustrations (sometimes with difficulty: the typical clause-initial position of *briefly* as a disjunct, for example, is not given). Moreover, as in *LDOCE*, the prosodic differences are entirely ignored.

The functional differentiation illustrated by *briefly* and *frankly* is not an isolated phenomenon, but characteristic of a whole range of adverbs such as *simply*, *literally*, *personally*, *clearly*, *naturally*, *superficially*, *technically*, *ironically*, *happily*, *hopefully* (for a detailed description and classification of adverbs, see Greenbaum 1969, Allerton & Cruttenden 1974, 1976, 1978, and Quirk et al 1985: Chapter 8). What is important to realize about these adverbs is that, although their function is generally signalled both positionally and prosodically, the prosodic distinction is often the more important one. Disjuncts, for example, which may occur initially, medially and finally in a clause, never take the sole intonation focus in clause-final position (as adjuncts tend to do), ie they must either be prosodically separated or entirely deaccented. Hence, although adjuncts and disjuncts may occur in the same syntactic position, they are always prosodically distinct, as shown in the following examples (from Allerton & Cruttenden 1976:48):

(3) (a) Richard played NATURALLY■ (adjunct)

(b) Richard PLAYED■ NATURALLY■ (disjunct)

In other words, *naturally*, *briefly*, *frankly*, and similar adverbs can be regarded as 'homomorphs' (cf Quirk et al 1985:71), whose function can only be fully clarified by means of intonation.

However, prosody does not only serve to distinguish disjuncts and adjuncts, nor is its role confined to signalling a contrast in intonational grouping. With conjuncts and disjuncts the choice of nuclear tone may be equally important. Conjuncts, for example, often have distinctive tones of their own:

(4) (a) Richard has RESIGNED■ THOUGH■

(b) *Richard has RESIGNED■ THOUGH■

(5) (a) BESIDES■ he didn't want to DO it■

(b) ?BESIDES■ he didn't want to DO it■

The choice of tone may also distinguish a whole functional class, as in the case of content disjuncts expressing value judgment (*curiously*, *fortunately*, *ironically*, *surprisingly*, etc), which generally occur with a fall-rise tone:

(6) FORTUNATELY■ he didn't come EARLY■

It may also distinguish functional subclasses, as in the case of content disjuncts expressing likelihood, which have different prosodic tendencies depending on the degree of conviction they convey. Thus, likelihood disjuncts expressing certainty (*clearly*, *definitely*, *certainly*, *obviously*, *naturally*, *of course*, etc) generally have an 'assertive' falling tone, while those expressing some doubt (*presumably*, *apparently*, *conceivably*, *possibly*, *probably*, etc) are more common with a fall-rise (for finer distinctions, see Allerton & Cruttenden 1974:15f; see also Chapter 9):

(7) CLEARLY■ he can WIN the match■

(8) PRESUMABLY■ he can WIN the match■

There are many other adverb classes with similar distinctive prosodic tendencies, but the examples given will suffice to show the strong connection between function and intonation within the adverb category.

The grammatical and prosodic properties of adverbs demonstrate two lexicographically relevant points. First, the adverbs form an important but very heterogeneous category that deserves a much more detailed and systematic treatment than is currently provided in dictionaries. Second, the prosodic behaviour of many adverbs is related to their semantic or pragmatic function. In other words, their prosodic potential is an essential part of their 'meaning', just as the complementation or selectional restrictions of a verb is part of the meaning of the verb. Indeed, a comparison with the treatment of verbs in dictionaries like *LDOCE* and *COBUILD* is illuminating. Whereas both dictionaries make a detailed and useful classification of verbs, only *COBUILD* has attempted something comparable for adverbs. In this respect, recent grammars like Quirk et al (1985) are far ahead of contemporary dictionaries.

A reasonable demand on future dictionaries is thus a treatment of adverbs that approaches the delicate description that is generally provided for verbs. Improvements should be possible in three respects:

(a) a classification of adverbs according to their functional use as modifier or adverbial adjunct, subjunct, conjunct or disjunct (with further subdivisions where relevant), (b) a rough indication of their positional tendencies, and (c) a representation of the typical prosodic behaviour of adverbial homomorphs. The prosodic notation could be simple, but should include an indication of intonational grouping (tone-unit boundaries) and predominant nuclear tone(s).[3] For lack of an international standard, a simplified version of the prosodic system developed by Crystal & Quirk (1964) and Crystal (1969, 1975) might be a suitable model. This system has gained increased currency in recent years, both in descriptive works like Quirk et al (1985) and in linguistic research (especially that based on the London-Lund Corpus; see pp 47ff and Altenberg 1986 and forthcoming a). It is also widely used in advanced language teaching materials (eg Crystal & Davy 1969, 1975, Leech & Svartvik 1975).

## 6.3 Discourse items

Another aspect of speech that is inadequately treated in dictionaries is the use of various types of 'discourse items' that are either rare outside speech or used in a speech-specific way (for a discussion of these, see Chapter 5). I am not here thinking of stylistically marked words that are typical of speech as a

**Table 6:1.** Distribution of discourse items in a sample (ten conversations) from the London-Lund Corpus.

| TYPES | FREQUENCY |
|---|---|
| RESPONSES | 2237 |
| *yes* | 727 |
| *m(hm)* | 658 |
| *no* | 259 |
| *yea(h)* | 217 |
| *oh* | 170 |
| *quite* | 35 |
| *I see* | 29 |
| *that's right* | 22 |
| *ah* | 17 |
| *right* | 15 |
| others | 88 |
| HESITATORS | 1226 |
| *ə(:)h* | 767 |
| *ə(:)m* | 430 |
| *m* | 29 |
| SOFTENERS | 438 |
| *you know* | 212 |
| *you see* | 119 |
| *I mean* | 102 |
| others | 5 |
| INITIATORS | 401 |
| *well* | 365 |
| *now* | 35 |
| others | 11 |
| HEDGES | 95 |
| *sort of* | 82 |
| *sort of thing* | 10 |
| others | 3 |
| EXPLETIVES (*God*, *heavens*, etc) | 52 |
| THANKS | 21 |
| *thank you* | 17 |
| *thanks* | 4 |
| APOLOGIES | 19 |
| *sorry* | 12 |
| others | 7 |
| ATTENTION SIGNALS (*hey*, *look*) | 6 |
| RESPONSE ELICITORS (*eh*, *right*) | 6 |
| POLITENESS MARKERS (*please*) | 5 |
| ORDERS (eg *give over*) | 5 |
| OTHERS (GREETINGS, etc) | 5 |
| TOTAL | 4516 |

predominantly informal medium (like *bloke*, *telly*, *fag*, etc), but of items that have a basically interactive and pragmatic function and therefore mainly occur in conversation, such as responses (*yes*, *no*, *quite*, *I see*, etc), 'softeners' (*you know*, *you see*, etc), hedges (*sort of (thing)*, etc), initiators (*well*, *now*), apologies (*sorry*, *pardon*), thanks (*thanks*, *thank you*), attention signals (*hey*, *look*), politeness markers (*please*), and greetings (*good morning*). A list of such items in a 50,000-word sample (ten conversations) from the London-Lund Corpus is given in Table 6:1 (derived from Stenström's inventory of categories in Chapter 5). The list is not exhaustive (only items occurring ten times or more have been listed separately), and the classification and labels are tentative, but the table gives a rough idea of the relative frequency of the most important types occurring in the conversation of educated British speakers.

As Stenström has pointed out (pp 137ff), several characteristics set these discourse items apart from other word classes. They are difficult, and often impossible, to analyse in traditional grammatical terms (as belonging to a certain part of speech or realizing a certain syntactic function); they generally contribute little to the propositional content of an utterance but rather fulfill various pragmatic functions in discourse; many of them take the form of (more or less) invariable multi-word units that are pointless to analyse internally.

The mere frequency of these items in everyday conversation is a sufficient reason why they should be given special attention by lexicographers. In the grammatically analysed sample of the London-Lund Corpus on which Table 6:1 is based, they account for 9.4% of all word-class tokens (see Table 6:2). What is more, discourse items constitute the fourth largest word-class category, outranked only by verbs, pronouns and nouns, but outranking such basic grammatical categories as prepositions, adverbs, determiners, conjunctions and adjectives. This means that, if frequency of occurrence is anything to go by in the compilation and organization of dictionaries, discourse items deserve to be treated with the same care and attention as the traditional word classes.

However, the decisive argument for treating discourse items as a category (or categories) of their own is not their frequency but their special discourse functions. Existing dictionaries, in so far as they include these items at all, run into obvious difficulties when they try to describe them under traditional word-class labels. Again, *LDOCE* and *COBUILD*, which probably give more attention to these speech-specific expressions than other dictionaries, may serve as examples.

**Table 6:2.** Relative frequency of major word classes in a sample (c 50 000 words) from the London-Lund Corpus.

| **Word class** | **%** |
|---|---|
| Verbs | 20.1 |
| Pronouns | 17.3 |
| Nouns | 14.3 |
| Discourse items | 9.4 |
| Prepositions | 9.2 |
| Adverbs | 9.0 |
| Determiners | 7.9 |
| Conjunctions | 6.3 |
| Adjectives | 6.0 |
| Predeterminers | 0.3 |
| Miscellaneous | 0.2 |

Let us start with *LDOCE*. Broadly speaking, *LDOCE* approaches these items in two ways: they are either (a) given a traditional word-class label and assigned a separate main entry or (b) presented in a subentry under a related but functionally different word. As a result, we find functionally similar or identical items treated under a number of different word-class labels and, conversely, functionally distinct items assigned to the same word class. I will give a few examples of what this traditional 'straitjacket' may lead to.

Symptomatically, a favourite word-class label for many discourse items in *LDOCE* is 'interjection', which is used for such functionally distinct words as *please* (politeness marker), *sorry* (apologizer), *well* (initiator), *hey* (attention signal), *good afternoon* (greeting), *thank you* (thanks), *damn* (expletive), and *oh* (which functions variously as response, initiator and exclamation in the corpus, in contrast to *eh* and *ah*, which are predominantly elicitor and response, respectively; cf Aijmer 1987). Another frequent word-class label is 'adverb', which is resorted to for a number of functionally disparate items such as *really* (elicitor, response), *now* (initiator) and *yes* (agreement).

An illustration of the reverse inconsistency - different word-class labels for the same discourse function - is equally illuminating. Thus we find, for example, response items variously classified as 'interjection' (*oh*, *ah*) or 'adverb' (in separate entries: *yes*, *yeah*, *OK*, *certainly*, *sure*, etc; in subentries under a related adverb: *no*, *quite*, *really*, *right*, *exactly*, *fine*, etc), or presented in a subentry under a pronoun (*that's it*), adjective (*that's right*, *right oh*) or verb (*I see*). Similarly, among expletives we find *gosh*, *(oh) dear* and *damn* as separate entries labelled 'interjection', but *dammit* (or rather *damn it*) under the verb *damn*, *God* under the noun *God*, *my* under the determiner *my*, and *for God's sake* under the noun *sake*; among hedges, *as it were* is treated in a subentry under the conjunction *as*, and *sort of* under the noun *sort*; among apologies, we find *sorry* in a separate entry labelled 'interjection', but *pardon* (with variants) in subentries under both the noun and the verb *pardon*; and among initiators, *well* is treated as an 'interjection' in a separate entry, but *now* in a subentry under the adverb *now*.

As these examples show, a traditional categorization of discourse items does not only result in functional inconsistency, but in functional inappropriateness. For example, in what way is *yes* or *OK* an 'adverb', and *sort of* (as in *sort of odd*) a 'noun'? Or *my* a 'determiner' and *good afternoon* an 'interjection'?

An additional complication is the fact that many discourse items are multi-word combinations. Such combinations are sometimes given a separate main entry, but more often presented (or merely illustrated) in a subentry under a related look-up word, which gives rise to further inconsistencies. Hence, we find *thank you* and *good afternoon* treated as interjections in separate entries, whereas the softeners *I mean*, *you see* and *you know* are treated under the verbs *mean*, *see* and *know* respectively, the smooth-over *never mind* under the verb *mind* (with a cross reference at *never*), the hedges *as it were* and *sort of* under the conjunction *as* and the noun *sort* respectively, and the responses *that's it* and *that's right* under the pronoun *it* and the adjective *right*. I am aware of the practical problems involved (lexical items must be easy to find in a dictionary), but practical convenience should not be allowed to, and indeed need not, affect the functional classification of dictionary items. Many types of multiword items are already systematically presented as separate entries in *LDOCE* (eg compounds and certain idioms) or listed separately at the end of a main entry (phrasal verbs); a similar recognition of multiword discourse items as independent units or parts of speech would therefore reduce also this type of inconsistency.

By contrast, *COBUILD* has managed to avoid most of these problems by two types of innovation. First, the dictionary entries are not primarily organized in terms of grammatical word class but on the basis of other criteria, such as frequency of use, independence of meaning and concreteness. Word-class information is instead given in abbreviated form in the extra marginal column provided in *COBUILD* (eg 'N COUNT' or 'ADV SEN'). This arrangement has the advantage that the grammatical status of an item does not determine the organization of an entry and, hence, has less serious consequences when the grammatical classification is inadequate.

The second innovation is of greater theoretical interest. The compilers of *COBUILD* have recognized the deviant, 'asyntactic' nature of many discourse items and introduced two new functional labels, 'CONVENTION' and 'PHR', to capture this. The first is applied to standard expressions with 'established form and meaning' that can occur as single independent utterances (*s v* CONVENTION, p 310) and is typically used for interactive expressions of various kinds, such as responses, greetings, attentions signals, apologies and thanks. The second label, 'PHR', is applied to multiword expressions that are more or less invariable and display some degree of integration in clause structure; when such a phrase has a typical clause function (as adverbial, object, etc), or occurs in a particular syntactic pattern, this is indicated after the category label. Discourse categories covered by this label are hedges (eg *as it were* = 'PHR: USED AS ADV SEN'; *sort of* = 'PHR: USU + ADJ/PAST PART' or 'PHR: USU + VB/ADV/PREP'), softeners (eg *you see* and *I mean* = 'PHR: USED AS ADV SEN') and the smooth-over *never mind* (= 'PHR: ONLY IMPER, IF + PREP THEN *about*').

In addition to these two labels, *COBUILD* uses two more conventional word-class symbols, 'EXCLAM' (corresponding to the traditional term 'interjection') and 'ADV SEN' ('sentence adverb(ial)') to describe various discourse items.

Thus, while *LDOCE* scatters discourse items over nine traditional categories, most of which are functionally inappropriate or misleading, *COBUILD* has reduced these to four: 'CONVENTION', 'PHR', 'EXCLAM' and 'ADV SEN'. Of these, 'CONVENTION' is typically used for interactive expressions (Responses, Greetings, Attention signals, Apologies and Thanks), 'EXCLAM' for Expletives, 'PHR' for Hedges, Softeners and Smooth-overs, and 'ADV SEN' for Initiators and the Politeness marker *please*. The result is a broad classification that makes no attempt at distinguishing any of the finer discourse categories recognized in Table 6:1 (such distinctions are generally, as in *LDOCE*, made informally inside each entry). In other words, the classification is not primarily discourse-oriented but grammatical: it recognizes such features as syntactic independence ('CONVENTION', 'EXCLAM'), syntactic complexity and degree of 'frozenness' ('PHR'), and syntactic function ('ADV SEN').[4] Yet, because it is systematic and avoids forcing speech-specific items into a traditional straitjacket, it is more satisfactory than *LDOCE*'s approach.

This does not mean that the *COBUILD* system is without inconsistencies. Responses, for example, are generally classified as 'CONVENTION', but *oh* and *good* are labelled 'EXCLAM' and *certainly* 'ADV SEN' in entries illustrating clear response functions. For *ah* no label is given. Expletives are normally classified as 'EXCLAM', but *for God's sake* is coded 'CONVENTION'. Softeners are labelled 'PHR:AD SEN', but *you know*, the most frequent type, has no label. Initiators are described as 'ADV SEN', but *well* is presented without a label. *Never mind* is classified as 'PHR', but since it is frequently used as an independent utterance, 'CONVENTION' would be an equally valid alternative.

Other interesting features to examine are the coverage, descriptive delicacy, and relative prominence that the two dictionaries give to discourse items. Since both dictionaries claim to describe natural English as reflected in extensive language corpora, we might expect them to be fairly exhaustive in their treatment of discourse items. Superficially at least, this also seems to be the case. Both include most of the discourse expressions listed in Table 6:1 in one way or another, and direct omissions are rare. Both fail to include the support *m(hm)*, which might be dismissed as a marginal lexical item but is in fact the second most frequent affirmative response in the London-Lund Corpus and an indispensable ingredient in any natural conversation (cf Tottie 1989). The common hedge *sort of thing* (see Aijmer 1984 and 1986) is also absent in both dictionaries (although there is a misleading cross-reference from *sort* to *thing* in *COBUILD*). *LDOCE* ignores *look* as an attention signal and merely mentions *I mean* in a usage note. *COBUILD* has omitted *my* in its exclamatory use and fails to give attention to *that's right* and *right*, which are both frequent responses in the corpus (*right* is also common as a response elicitor; cf Stenström 1987).

If the coverage of discourse items is about the same in the two dictionaries, the treatment of the included items differs considerably. This is partly due to the space available (*COBUILD* is the larger dictionary), but also to a difference in editorial policy. In accordance with its 'single look-up' principle, *COBUILD* presents all relevant information about an item in a single (sub)entry, whereas *LDOCE* may counterbalance a brief entry with more extensive stylistic or pragmatic information in separate usage notes or full-page language notes (eg on 'Addressing people', 'Apologies', 'Invitation and offers', 'Politeness', 'Thanks', etc). *LDOCE* also has a greater tendency to present discourse items as separate entries, while *COBUILD* describes them in numbered subsections or paragraphs of a main headword (a consequence of *LDOCE*'s policy to organize entries according to word class but *COBUILD* according to word form). However, if we disregard these differences (and the pedagogical merits they may have), *COBUILD*'s treatment is generally more systematic, detailed and exhaustive than *LDOCE*'s. This is revealed in several ways. *COBUILD* is more consistent in recognizing the multifunctional character of many discourse items, and consequently takes greater care to describe and illustrate each function. For example, while *COBUILD* identifies about a dozen different uses of *yes*, *no* and *well*, and half a dozen uses of *ah*, *now* and *please*, *LDOCE* merely gives two or three. Moreover, in *COBUILD* each function is clearly set off in numbered subsections and highlighted in bold face, while *LDOCE* often lumps together several uses in a single (sub)entry or confines itself to a casual illustration in passing (as in the case of *I see* and *that's right*), a tendency that is not always redeemed by its usage and language notes.

Since frequency of use is one of the organizing principles in *COBUILD*, one might expect discourse items to occupy an early position in its entries. This is also generally the case: discourse items like *well*, *you know*, *you see*, *I mean* and *sort of* generally crop up earlier in *COBUILD* than in *LDOCE*. But there are exceptions: the initiator *now* is presented as the third sense of *now* in *LDOCE* but only as the tenth in *COBUILD*, and the frequent response *I see* is illustrated (though casually) much earlier in *LDOCE* (sense 5) than in *COBUILD* (sense 10). Indeed, sometimes one wonders what impact the frequency principle has been allowed to have on the ordering of subentries in *COBUILD*. To take just two examples, *you know* and *you see*, which are probably the two most common discourse expressions in English (cf Sinclair & Renouf 1988:151f), only appear as senses 15 and 18 of *know* and *see* respectively (in *LDOCE*, it should be added, they are given even more insignificant positions).

To sum up, both dictionaries include most of the items listed in Table 6:1, but their treatment of them differs. While *LDOCE* pays special attention to the pragmatic use of discourse items in separate 'Usage and Language Notes', *COBUILD* tends to give greater prominence to the items in the dictionary entries themselves (a feature that is partly determined by its 'single look-up' policy, partly by its greater attention to corpus frequency). But what is more important, *COBUILD* is more felicitous in its grammatical classification of discourse items. Although the classification is crude in some respects, it avoids the inadequacy that is inevitable with a traditional system.

## 6.4 Conclusion

Speech differs from writing in many fundamental ways. I have here touched on two speech-specific phenomena, the use of intonation to differentiate adverbial functions, and the use of lexical items with pragmatic functions that are difficult to describe in traditional grammatical terms. If we wish dictionaries to reflect the spoken language (which they surely should do), they must also recognize these phenomena and find methods of representing them in an adequate way. As this scrutiny has shown, even recent 'speech-oriented' dictionaries often fail to do this, although improvements are noticeable in some respects. The suggestions I have made here include

(a) a functionally relevant classification of adverbs,
(b) a systematic treatment of discourse items as linguistic categories of their own, and
(c) the introduction of a simple prosodic notation (indicating tone-unit boundaries and major nuclear tones) to clarify functional differences in illustrations of speech.

In all these areas, natural speech corpora (such as the London-Lund Corpus) provide a rich source of information that is likely to be of great benefit to lexicographers and dictionary users in the future.

## Notes

1 This chapter is a revised version of a paper (Altenberg 1988) presented at the conference on 'Standardization in Computerized Lexicography', Saarbrücken, 15-17 October 1986.

2 There are important differences in the way the two dictionaries have made use of their source corpora. While *LDOCE*'s corpus (the Longman Citation Corpus) mainly seems to have served 'as a *basis* for creating natural examples' (Summers 1988:13), the *COBUILD* database (the Birmingham Collection of English Text) was used systematically as textual evidence for the compilation of the dictionary, not only in the selection of examples (phrases, collocations, etc) but in determining word meanings and estimating their relative importance (see Sinclair 1985 and 1987a). Moreover, it is unclear how much natural speech was included in the Longman Corpus. For an account of the *COBUILD* Corpus, see Renouf (1987).

3 The only dictionary I know of where an attempt has been made to use intonation to elucidate functional differences is the *Oxford dictionary of current idiomatic English*. An illustration of this is its treatment of *you know*, which is given four different entries (Vol 2, p 603f), each with its functional and intonational specification (eg *you know* 1: 'you know or understand very well [...] often preceded by short pause; fall-rise tone on *know*.'). One may disagree with details in these explanations, but the approach is laudable and worthy of imitation.

4 A more accurate way of describing *COBUILD*'s classification is to say that grammatical criteria have been used as far as they have been applicable ('ADV SEN', 'PHR'). Expressions that have not been possible to define grammatically in a simple way have been divided into two main classes, 'EXCLAM' and 'CONVENTION'. Of these, the latter can be regarded as a broad residual category consisting of a range of predominantly interactive but functionally distinct subclasses. Thus, although an 'interactive' category is well worth recognizing, what is less satisfactory about it is its discourse-functional diversity. To conflate so many distinct subtypes under one label is comparable to recognizing only one category of verbs or one category of pronouns. In addition, the name 'CONVENTION' is unfortunate, since conventionality of expression is not unique to this category.

# 7

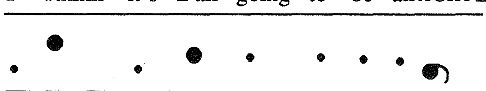
# Some functions of the booster

*Bengt Altenberg*

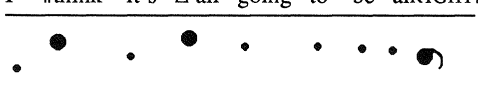
## 7.1 Introduction

In the prosodic system used for the transcription of the London-Lund Corpus, a step-up in pitch in the intonation contour of a tone unit is called a 'booster'.[1] The booster system comprises three variants illustrated by *all* in the following examples (from Crystal 1969:146):

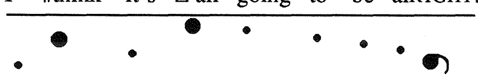
(a) I ||think it's ▵ all going to be alR↘IGHT■



(b) I ||think it's △ all going to be alR↘IGHT■



(c) I ||think it's Δ all going to be alR↘IGHT■



In (a), *all* has a slightly higher pitch than the preceding (stressed or unstressed) syllable; in (b) it has a higher pitch than the next previous pitch-prominent step-up syllable (in this case the onset, but alternatively any preceding syllable marked by continuance, booster or high booster); in (c) it has a pitch that is very much higher than the next previous pitch-prominent syllable. These three degrees are referred to as 'ordinary', 'high' and 'extra-high' booster, and are here marked by a small (▵), medium-sized (△), and big (Δ) triangle, respectively.

Although the booster system is merely part of a larger system of pitch-range contrasts in the intonation contour (see Crystal 1969:144ff), it contributes in an essential way to our impression of melody and variation in speech. Without it, continuous speech would tend to sound monotonous and lifeless, consisting primarily of a succession of falling contours (which are sometimes said to represent the 'unmarked norm' in English, although this norm may in fact be restricted to mechanical forms of reading; see Crystal 1969:232, Cruttenden 1986:127) and varying mainly in terms of onset selection and nuclear tone movement.

Generally speaking, any pitch obtrusion (whether up or down) gives prominence to a word, but a step-up from a predominantly falling contour seems to have a special foregrounding or heightening effect. However, little has been done to explore the functional aspects of the booster. Crystal, for example, who has provided the most detailed account of the booster and its role in the prosodic organization of the 'head' of the tone unit, only touches on its function in very general terms (1969:225-233): it is said 'to spread relative prominence over the words in the head, and to add prosodic variety to connected speech' (p 233). It is also shown to have a strong correlation with various types of emotional excitement (pp 301-305). Cruttenden (1986:88) notes that a 'pre-nuclear pitch accent' may be used to indicate the beginning of the focus in a tone unit (the end being marked by the nucleus).

The most ambitious attempt to specify the function of pitch obtrusions in the intonation contour is probably Bolinger's discussion of the 'hat pattern' (1986:46ff) and what he calls Profiles A and B (both of which begin with a step-up in pitch; see pp 142ff). Although Bolinger's prosodic system differs in many respects from that of the London-Lund Corpus (it lacks, for example, the structural organization of tone units into prehead, head, nucleus and tail, and there is consequently no one-to-one relationship between his upward-jumping pitch accents and the booster), it is interesting to note that Bolinger associates the initial step-up of the A and B Profiles (or any hat-like contour that begins with these profiles) with an 'annunciatory' or 'thematic' function, which contrasts with the 'terminatory' or 'rhematic' function of a final (nuclear) accent. In particular, while the A Profile (a step-up followed by a drop) is said to be the assertive profile par excellence (p 164) which 'singles things out' (sometimes contrastively), the B Profile (a step-up not followed by a drop) is said to be 'connective' and associated with compounds and other close-knit expressions. Moreover, the B Profile is used not so much to inform as to 'enhance' and is consequently frequent with quantifiers and affective modifiers (pp 169ff):

```
           áll of them        every ó
I've got               fín           n
                                      e.
                          ished,

         rótten
You're a        lí
                  ar.
```

Thus, although there have been some attempts to specify the functions of the booster (or booster-like phenomena), these have generally been vague and impressionistic and partly in conflict with each other. The purpose of the present study is to make a preliminary survey of the functions of the booster in a small sample from the London-Lund Corpus. The sample consists of five texts: four surreptitiously recorded conversations (texts S.1.4, S.1.5, S.1.6 and S.1.9) and one prepared monologue (text S.12.6, a public but rather informal lecture), each text totalling some 5000 words. The sample was originally chosen to reveal possible differences between speakers and speech varieties, but I will not emphasize this aspect here.

Two important restrictions on the study should be mentioned. In the functional analysis, no distinction has been made between different degrees of booster (ordinary, high and extra-high), although such a distinction is no doubt both relevant and interesting. Moreover, I have concentrated entirely on boosters in the 'head' of the tone unit (ie between the onset and the nucleus), thus excluding boosters affecting a nuclear syllable. The reason for this is that one, rather trivial, function of the booster seems to be to prepare the way for a falling nuclear tone when the speaker has reached the baseline of his pitch range at the end of a tone unit (see Altenberg 1987a:33). A concentration on independent (non-nuclear) boosters thus increases our chances of isolating 'pure' booster functions unaffected by the requirements of nuclear tone direction.

My approach has been determined by the necessity to start 'from scratch' and to discover patterns of cooccurrence at different levels of linguistic description. Thus the booster will be examined for possible correlations with the following categories: (a) word forms, (b) word classes, (c) semantic categories, and (d) the information structure of the tone unit. The grammatical framework will be that of Quirk et al (1985).

## 7.2 General booster distribution

On average, every second tone unit in the material contains a booster of some kind, but only one in eight has a booster in the head of the intonation contour - the position that is our concern here. The use of a booster between the onset and the nucleus obviously requires tone units of some length. It is significant that while the average tone unit length in the texts is 4.1 words, tone units with a booster have an average length of 7.1 words. Moreover, boosters seldom occur in tone units shorter than four words.

Normally, there is only room for one booster (86%), but longer tone units may have two (12%) and exceptionally three (1%) or even four (0.4%) as in the following examples (where the relevant booster-marked words are italicized):

(1) and "‖THÂT was■ an EX‖{TRÈMELY} ABS▵TRÛSE 'talk■ [...] and • ‖on a ▵*topic* that ▵*most* people knew ▵*nothing* *what*▵*ever* A∆BÒUT■ (S.1.6:1003-1008)

(2) but ‖on the ▵*other* ▵*hand* you ▵*do* ▵*meet* • ▵SĔCRETARIES■ (S.1.5:1192)

For the same reason, when a booster occurs in the head of a tone unit, it tends to appear fairly soon after the onset, the average distance being 2-3 words.

High boosters are more common than ordinary boosters (57% and 42% respectively), while the extra-high variant is rare (1%) and mainly reserved for strong emphasis:

(3) and I ‖said it would be ∆*far* "BÈTTER■ ‖if the "▵mansion was ▵KÉPT■ (S.12.6:760-761)

(4) ‖this is the ∆*only* thing I've 'brought A∆WÀY from that LÉCTURE■ (S.1.6:944)

Depending on the type of booster occurring in the head of the tone unit (ie disregarding nuclear boosters), we can make a broad distinction between falling and rising heads (for a more detailed classification, see Crystal 1969:229ff). A head is falling if it contains no step-up higher than an ordinary booster, and rising if it contains at least one high booster. As shown in Table

**Table 7:1.** Booster distribution and types of head

| NUMBER OF BOOSTERS IN HEAD | FALLING HEAD | RISING HEAD | TOTAL | % |
|---|---|---|---|---|
| 0 | 5167 | - | 5167 | 87 |
| 1 | 254 | 401 | 655 | 11 |
| 2 | 24 | 70 | 94 | 1 |
| 3 | 3 | 6 | 9 | 0 |
| 4 | 0 | 3 | 3 | 0 |
| TOTAL | 5448 | 480 | 5928 | 100 |
| % | 92 | 8 | 100 | |

7:1, the great majority of the tone units in the texts have a falling head (92%), which is a natural consequence of the fact that most heads have no booster at all. Heads with a booster, on the other hand, are generally rising (63%), owing to the predominance of high boosters mentioned above.

## 7.3 The booster potential of words and word classes

After these preliminaries, let us now examine the functions of the booster. As a first step, it may be instructive to look briefly at its tendency to cooccur with different word forms in the two texts. Out of a total of 365 booster-marked words, those most frequently carrying a booster are the following:

| | | | | | |
|---|---|---|---|---|---|
| *very* | 24 | *is* | 7 | *can't* | 5 |
| *don't* | 12 | *quite* | 7 | *get* | 5 |
| *I* | 12 | *rather* | 7 | *going* | 5 |
| *just* | 11 | *think* | 7 | *only* | 5 |
| *one* | 9 | *read* | 6 | *other* | 5 |
| *this* | 9 | *that* | 6 | *really* | 5 |
| *what* | 9 | *three* | 6 | *same* | 5 |
| *all* | 8 | *always* | 5 | *some* | 5 |
| *he* | 8 | *any* | 5 | *Stoke* | 5 |

The list is headed by the intensifier *very*, which suggests that one function of the booster is to emphasize expressions of degree (the adverbs *just*, *quite* and *rather* further down the list are other intensifying words). The booster also seems to be associated with expressions of quantity (*one*, *all*, *three*) and reference (*I*, *this*, *what*, *he*, *that*, *other*, *same*), or a combination of these (*any*, *some*), as well as with denying, affirming or qualifying truth value (*don't*, *can't*, *is*, *really*, *think*).

However, apart from such vague indications little can be concluded from this list. Indeed, the most striking feature of the booster seems to be its wide lexical distribution. With the possible exception of *very* (which has a booster in nearly half of its occurrences), no words can be said to have a distinctive 'booster potential'. It is significant that the twenty-seven listed words merely represent 5% of all the booster-marked word types in the material (though 23% of the tokens); an additional 131 words (26%) attract a booster 2-4 times, but the great majority (80%) have a booster only once. In other words, the lexical patterning of the booster is too varied to be functionally revealing.

A more profitable approach may be to look at the booster potential of different classes of words (rather than individual word forms). If all word classes (however defined) had the same booster potential, the boosters in the material would be proportionately distributed over the classes. This is not the case, however: some classes tend to attract a booster more readily than others. Table 7:2 ranks twenty different word categories according to their booster potential, expressed in terms of a difference coefficient. The coefficient indicates the deviation of the recorded booster frequency for each category from its 'expected' (word-class proportional) frequency.[2] The coefficient may vary from +1 to -1; a plus value indicates overrepresentation of boosters in the category (high booster potential), a minus value underrepresentation (low booster potential).

As shown in the table, the word categories are differentiated along a gradient, with those having a high booster potential at the top and those having a low booster potential at the bottom; categories near the middle of the scale (roughly within ± 0.1) can be described as 'neutral'. Outstanding at the top of the scale we recognize the intensifier *very* and the class of predeterminers, chiefly represented by the quantifier *all* and the intensifier *quite*. Other quantifying classes also appear high up in the scale, such as postdeterminers (which include the cardinals), quantifying central determiners (*some*, *any*, *no*, *every*, *another*), and compound pronouns (which include the universal pronouns *anybody* and *anything*). We also find the archetypical 'open' word classes (*ly*-adverbs, adjectives, full verbs and nouns) in the top half of the scale, a fact which seems to link the booster with lexical salience. Moreover, the position of the modal verbs and *do* above the middle of the scale supports our previous observation that the booster seems to be associated with expressions of truth value.

**Table 7:2.** Booster potential of major word categories

| WORD CLASS | BOOSTER FREQUENCY | | DIFF. COEFF. |
|---|---|---|---|
| | RECORDED | EXPECTED | |
| *very* | 24 | 3.9 | 0.72 |
| Predeterminers (*all*, *quite*, etc) | 13 | 2.7 | 0.66 |
| Postdeterminers (*one, first*, *next*, etc) | 66 | 19.1 | 0.55 |
| Quantifying determiners (*some*, etc) | 11 | 3.5 | 0.52 |
| *ly*-adverbs | 32 | 10.4 | 0.51 |
| Adjectives | 96 | 34.0 | 0.48 |
| Indefinite compound pronouns | 11 | 4.1 | 0.46 |
| Relative pronouns | 7 | 3.0 | 0.40 |
| Full verbs | 182 | 97.3 | 0.30 |
| Modal verbs | 31 | 16.8 | 0.30 |
| *do* (all functions) | 10 | 12.0 | 0.18 |
| Nouns | 175 | 126.5 | 0.16 |
| Other adverbs | 88 | 64.8 | 0.15 |
| Demonstrative determiners | 9 | 7.8 | 0.07 |
| Demonstrative pronouns | 9 | 9.9 | - 0.05 |
| Subordinators | 11 | 13.8 | - 0.11 |
| *be* and *have* (all functions) | 23 | 47.0 | - 0.34 |
| Prepositions | 23 | 66.6 | - 0.49 |
| Other pronouns | 8 | 39.4 | - 0.53 |
| Other determiners | 6 | 58.1 | - 0.81 |
| Infinitive marker *to* | 1 | 14.5 | - 0.87 |

Thus, certain booster functions suggested in our original word list seem to be confirmed in Table 7:2. In other respects, however, the scale displayed in Table 7:2 is rather disappointing. Most of the word classes are functionally opaque, and the scale as such does not permit any conclusions beyond the fact (though interesting enough) that the booster is unevenly distributed across the word classes. The most we can say is that classes containing lexical (or content) words tend to have a higher booster potential than those containing grammatical (or function) words, but even this generalization is weakened by the existence of many grammatical classes at the top of the booster scale.

There are several reasons why the word categorization in Table 7:2 fails to reveal anything interesting about the booster (cf Altenberg 1987a:133ff). Many classes are functionally heterogeneous, eg the adverbs and postdeterminers which include subcategories with diverse functions. Conversely, many functions cut across the word classification, as is illustrated by the 'quantifying' function which may be realized by several grammatical categories (predeterminers, determiners, pronouns, nouns, adverbs). Hence, if we wish to determine the uses of the booster, we must go beyond the word classes and examine their functions from other perspectives.

## 7.4 Some functions of the booster

The booster functions that emerge from the material can be classified in various ways and at different levels of delicacy. In the following survey, I have used a rather broad semantic-pragmatic classification, but I am aware that finer distinctions are possible. Needless to say, there is a great deal of overlap between the different functions, and it often happens that several functions cooccur in the same example. However, I shall disregard these classificational problems here and describe each function as if it were a clear-cut and unproblematic category. For each category I will indicate some typical lexical and grammatical correlates of the booster.

### 7.4.1 Intensifying quality and quantity

If we want to express a high or low degree of something we use an intensifying word, an amplifier (scaling upwards) or downtoner (scaling downwards). A typical use of the booster is to give prosodic prominence to such degree words, especially amplifying adverbs like *absolutely*, *completely*, *extremely*, *jolly*, *perfectly*, *quite*, *right*, *so*, *terribly*, *too*, *utterly* and *very*:

(5) you ||just SÀT■ and ||had a Δ*jolly* 'good "▵{GÌGGLE} at the ▵things he was ΔSǍYING■ (S.1.6:773-774)

(6) ||and it's Δ*very* ÌNTERESTING■ (S.12.6:702)

(7) you ||go • Δ*right* 'down • the 'main RÓAD■ (S.12.6:831)

Booster-marked downtoners (*a bit*, *about*, *a little*, *rather*, *quite*, *somewhat*, *virtually*, etc) are less common:

(8) I [m m] pre||sented • «a» ▵*rather* ABSǓRD re'port in a 'way■ (S.1.4:889)

(9) ‖that's *a*△*bout* WHAT■ ‖forty [f] • ‖over 'fifty per ▵CENT■ (S.1.4:734-735)

(10) ‖[?]I 'think 'Malcolm's 'TWENTY-SEVEN■ ‖TWENTY-EIGHT■ - per‖haps *a* △*bit* MORE■ (S.1.6:46-48)

Boosters are also frequent with amplifying adjectives, especially those denoting extreme degree, such as superlatives and what might be called 'inherent superlatives':

(11) ‖he's an △*absolute* ▵poppet HIM▵SELF■ (S.1.6:1065)

(12) ‖and △he MARRIED■ ‖one of the ▵*leading* ▵ACTRESSES of the TIME■ (S.12.6:512-513)

(13) he ‖gave a △*huge* △FEAST■ (S.12.6:523)

(14) ‖he 'gave a *ter*"▵*rific* 'lot of WORK■ (S.12.6:614)

Intensifiers typically modify gradable qualities and quantities. A booster cooccurring with an intensifier can consequently be said to function as a prosodic intensifier. Often, however, we wish to emphasize not the degree of a phenomenon but a special aspect of it or its scope of reference. Restricted reference is typically expressed by focusing subjuncts (eg *at least*, *only*, *just*, *wholly*) or adjectives (*exact*, *main*, *peculiar*, *special*, *unique*), and is often prosodically emphasized by a booster:

(15) «it» ‖being a [l] ▵*wholly* ▵ladies' PARTY■ (S.12.6:1001)

(16) ‖«even» △{READING} "△*just* 'those BITS■ (S.1.4:283)

(17) ‖it's a △*special* △MEETING 'he's ADDRESSING■ (S.1.4:1105)

(18) ‖now △this was a ▵*u*'*nique* THING■ (S.12.6:929)

The opposite extreme - wide scope of reference - is mainly expressed by adjectives denoting totality (*complete*, *universal*, *whole*, etc). These also tend to attract a booster:

(19) I sup‖pose 'this is the [kəmpl] - *com*▵*plete* △CHOICE■ (S.1.4:79)

(20) but ‖this was de'feating [ði:] ▵*whole* OBJECT of it■ (S.1.4:1166)

(21) (...) as ‖though it's a kind of △*communal* "△LINE on 'this■ (S.1.4:667)

The boosters illustrated so far have accompanied words expressing degree (high/low) or scope (wide/narrow). In these cases prosody can be said to support lexis: an idea already expressed in lexical terms is reinforced by a booster. However, words do not have to indicate an extreme quality to be emphasized by a booster. It is often sufficient that they convey something that the speaker thinks is worth drawing attention to, generally something attractive or peculiar (*beautiful*, *lovely*, *famous*, *good*, *great*, *interesting*, *new*, *old*, *rough*, *short*, etc):

(22) (...) to "‖SHÒW■ that ‖this •‖would be an △*interesting* COMPÀRISON■ (S.1.4:348-350)

(23) so you'll ‖never 'be in that △*happy* PO△SÌTION■ (S.1.5:334)

(24) [it] ‖had CON△NÈCTION■ ‖with our △*famous* Thomas △GRÁY■ (S.12.6:700-701)

(25) ‖one of the RÉASONS■ ‖why our △*lovely* "CHÚRCH■ (...) ‖is in such a △*good* 'state of preser'vation TO"DÁY■ • ‖is (...) (S.12.6:1058-1061)

Qualities of this kind are generally gradable and therefore, as we have seen, often intensified by a degree adverb. When they are not, as in (22-25), a booster may achieve much the same effect - compare for instance examples (6) and (22). Consequently, in its intensifying function the booster can either reinforce an existing intensifier, or replace it. In this respect speech has a significant advantage over writing: speakers may not only reinforce a lexical expression intonationally, they can also use intonation to emphasize something that is not, or cannot, be expressed verbally. This advantage is highlighted in cases where a booster accompanies a nongradable adjective, as in (11-14) above. Such adjectives can only be intensified by prosodic means.

This independent intensifying function of the booster is especially useful with expressions of quantity. Although some quantifying expressions can of course be intensified lexically (eg *very many*, *many more*, *too much*; cf also *about fifty per cent* and *a terrific lot of work* above), many cannot, especially if they already denote extreme degree (as the 'totality' adjectives above: *communal*, *whole*, *complete*). In such cases a booster (or, sometimes, a maximizer like *absolutely* or minimizers like *at all*, *whatever*) may be the only means of emphasizing the size of the quantity. This is no doubt the reason why so many expressions of absolute (maximal or minimal) quantity or reference appear with a booster in the material (eg *all*, *every*, *no*, *any*, *most*, *anything*, *anybody*):

(26) we ‖met △*all* the A△MÈRICAN■ ‖CHÁIRMEN■ (S.12.6:793-794)

(27) they ‖had △*no* DISTRÀCTION■ (S.12.6:329)

(28) ‖and Δ*every* DÀY■ ‖we it was ‖my 'job to Δpump the water 'up (...) (S.12.6:162-163)

(29) ‖cos [?] Δ*any'body* with Δ*any* SÈNSE■ would ‖read the PLÀY in the 'trans•LÁTION■ (S.1.4:305-306)

However, the booster is not only a useful means of emphasizing extreme quantities. Any nongradable quantity (whether great or small) that the speaker judges to be in some sense extreme or special in the context may be highlighted by a booster:

(30) ‖but in 'this Δthere are Δ*two* PÒINTS■ (S.1.4:671)

(31) well ‖this CÒLLEGE■ has ‖been re'sponsible for the 'loss of Δ*three* • LÌCENCES■ (S.1.9:497-498)

(32) we ‖ÙSED to■ • ‖fetch the ΔMÌLK■ • ‖from 'Grange FÁRM■ ‖which was a'bout a Δ*quarter* of a MÌLE■ ‖down the ΔRÒAD■ (S.12.6:123-126)

(33) ‖and perΔhaps Δ*some* of 'you • could REΔMÈMBER■ what ‖hot 'numbers they 'really WÈRE■ (S.12.6:230-231)

### 7.4.2 Emphasizing truth and modality

In the cases illustrated so far the booster has chiefly served to intensify or emphasize an element within the noun phrase - the typical domain of notions like quality, quantity and reference. States and events can also be intensified (if they are gradable), but it is more common that they are evaluated with regard to their truth value or likelihood. Thus, a speaker may assert or deny the truth value of an utterance, or he may take a middle ground and express some doubt or value judgement about its content. Such 'modal' meanings are typically expressed by verbs and attitudinal adverbs interacting with the negative particle *not*. In all these cases the booster is common as a prosodic emphasizer.

The truth value of an utterance may be reinforced by content disjuncts and emphasizing subjuncts like *certainly*, *really*, *indeed* and *of course*. These are often supported by a booster:

(34) ‖and [ə] • he Δ*certainly* Δstirred the Δ{PLÀCE} ΔÙP■ (S.1.6:706)

(35) ‖well they "Δ*really* 'haven't "ΔÀNY {‖RÈASON to■ }■ (S.1.4:839)

(36) ‖it was *in*Δ*deed* a Δday ÒUT for 'us■ (S.12.6:113)

But just as we have seen that a booster can replace an intensifying adverb, it can also function as an independent emphasizer of truth value, whether this is affirmed or denied. Indeed, this is one of the most frequent uses of the booster in the material:

(37) ‖he said Δ*is* the 'Six BELLS■ (...) ‖still THERE■ (S.12.6:687-690)

(38) «but ‖would» you Δ*get* Δthat ΔUP 'any'where■ (S.1.4:531)

(39) well ‖I Δ*don't* ΔKNOW■ (S.1.4:1051)

Modality often involves various expressions of possibility, likelihood and doubt. These notions may be realized in various ways, eg by content disjuncts (*probably*, *perhaps*, etc), by verbs of 'hedging' (*presume*, *suppose*, *think*, etc), and by modal auxiliaries (*can*, *may*, *must*, etc), all of which are frequently emphasized by a booster:

(40) ‖or *perΔhaps* it ΔIS 'lung 'cancer■ (S.1.4:1043)

(41) ‖so I *preΔsume* it 'is for ▵any'body in the ▵faculty of ΔARTS■ (S.1.4:1141)

(42) I ‖«said» she Δ*might* FAIL■ (S.1.4:891)

### 7.4.3 Emphasizing contrasts

Another frequent use of the booster is to highlight contrastive elements, as shown in the following examples:

(43) B ‖are you Δdoing ▻two or ONE • ▻paper this ▻year■ •

A ‖only ▵ONE■ •

B ‖YES■ • but ‖that's a 'main "ΔLINE 'paper■ ‖ISN'T 'it■ so ‖probably [j] ▵*you* will 'have "▵*more* SCRIPTS■ than ‖I shall 'have in ▵*two* Δ*special* ΔSUBJECTS■

(S.1.4:860-866)

(44) ‖I had a 'seminar TODAY■ «in which» • ‖people Δ*hadn't* 'read the «ΔSTUFF»■ (...) ‖I ΔSAID■ ‖shall we [?] 'do 'something ELSE■ or ‖shall I Δ*tell* you a'bout the 'plans for the ▵*new* SYLLABUS■ but "‖HALF of them had READ it■ and the ‖others 'said they 'wanted ▵*me* to ΔTALK a'bout it■ ‖so I {ΔDID} for a BIT■ (...) but we're ‖going to go 'on with it ▵*next* [ə]- TIME■ (S.1.4:1081-1099)

As these examples demonstrate, the role of contrastive highlighting is partly shared by the onset and the nucleus. This possibility is available when the contrastive element occurs within the normal domain of these two features, ie at the beginning and end of the tone unit, as illustrated by for example *else*, *half* and *others* in (44). (The nucleus may of course also have a booster of its own and be fronted for extra emphasis.) There is thus an interplay between the booster and the other markers of pitch prominence in the tone unit: words requiring special highlighting can be seen as competing for prosodic attention, and what type of prominence they get is partly determined by the degree of importance the speaker attaches to them and partly by their position in the tone unit.

In the contrastive category we may also include cases of deictic emphasis, which often involve a contrast, implicitly or explicitly. These are typically realized by demonstrative pronouns and determiners, and often emphasized by a booster, as illustrated in

> ||*are* Δ*these* S↗IMILAR
> (pointing at two pictures in a collection), or
>
> ||*that was* Δ*this* T↘ERM
> (implying a contrast with 'last term').

### 7.4.4 Grouping function

In the cases presented so far, the booster has generally served to highlight a single word for various purposes (intensifying, emphatic or contrastive). But frequently the booster also has what might be called a 'grouping' function, signalling the beginning of a group of words that belong together semantically and syntactically. Since the end of such groups is normally in focus, the booster and the nucleus can be said to delimit the group prosodically and give it a 'unifying contour' within the tone unit. The types of construction most commonly highlighted in this way are compounds stressed on the last element, verb-particle combinations with a stressed particle, postmodified and coordinated noun phrases, and various name combinations:

(45) it was ||turned 'into a Δ*country* *CL↘UB* ■ (S.12.6:739)

(46) ||so we 'had to "Δ*spin* it ↘OUT ■ (S.1.4:1085)

(47) I ||brought Δ*cups* and S↗AUCERS ■ (S.1.4:924)

This grouping function is very frequent in the material (see also *a* ▵*day* ÒUT and ▵*faculty of* ∆ÀRTS in (36) and (41) above), and it is possible that it can be given a more general explanation. I will return to this possibility below.

### 7.4.5 Other functions

Apart from the functions mentioned, several other uses of the booster are suggested in the material. I will only mention a few of these here, however.

In some cases the booster seems to have the purely rhythmical function of supplying the expected degree of prominence in a patterned sequence:

(48) [the family] ‖RÁN■ as a com∆plete 'self-supporting ÙNIT■ – ‖butlers 'cooks ▵*servants* 'gamekeepers ▵*gardeners* and the LÒT■ (S.12.6:444-446)

(49) ‖Rifleman HÁRTLEY■ ‖number ▵*so* and so and SÓ and so■ [rərə] re‖port to ▻Colonel ▵J A GÌLLIATT■ (S.12.6:1225-1227)

In other cases (reminiscent of the contrastive use) it functions as a 'repair' signal, reinforcing a correction after a hesitation or false start:

(50) so ‖I «had to» ∆*did* it A▵GÀIN■ (S.1.4:292)

(51) we DE‖CÎDED■ or ‖rather ▵*it* was [ə] DE▵CÌDED■ to ‖pull it DÒWN■ (S.12.6:754-756)

In (51) the booster signals a change of clausal theme (in the sense of Halliday 1985b: 38ff). Since themes are by definition clause-initial and normally carry given information, they tend to be prosodically unmarked or, if salient enough, marked by the onset. However, if a theme occurs after the onset, as in (51), it may of course be highlighted by a booster like any other element that is felt to be of special importance. Such 'thematic highlighting' is quite common in the material (about 15% of the boosters affect a thematic element), especially when the theme is contrastive (as *you* in (43) and *it* in (51) above), but also when it introduces a new referent (52) or point of departure (53), or when an old referent needs to be reinforced (54):

(52) ‖Miles LABÒRATORIES■ (...) ‖bent over "∆BÀCKWARDS■ to do "‖everything PÓSSIBLE■ • ‖their *lib∆rarian* 'spent "▵YÊARS■ ‖looking up ▵all the DÒCUMENTS of GRÁY■ (S.12.6:818-823)

(53) ‖but the ∆*trouble* ÌS■ it was ‖so "ÌNTERESTING■ • to "‖DÒ■ (S.1.4:1162-1164)

(54) ||nice BOY■ – ||sure △*he'd* HELP you■ if you ||got STUCK■ (S.1.6:27-29)

Often a booster serves to mark a secondary (subordinate) theme in the tone unit:

(55) I ||don't know △*what* I'm 'going to 'do a'bout this △SEMINAR TOMORROW■ (S.1.4:952)

(56) «and» ||you △came a'way △AFTER 'WARDS■ and ||thought now △*what* have I ▷brought A△WAY■ (S.1.6:776-777)

(57) and ||then he says △*course* «if» you △don't UNDERSTAND this■ – this ||subject's △not for YOU■ (S.1.6:920-921)

In cases like these the booster acts much like a secondary onset, serving to announce the beginning of a new clause (55) or of direct speech (56-57).

## 7.5 Conclusion

This brief sketch of some recurrent booster functions in a sample of the London-Lund Corpus does not of course exhaust the uses of the booster in speech. Indeed, one striking feature of the booster is its functional versatility. Although certain (frequently overlapping) uses predominate in the examined material (see the rough estimation in Table 7:3), it is difficult to identify a

**Table 7:3.** Relative frequency of major booster functions

| FUNCTION | % |
|---|---|
| Truth and modality | 28 |
| Grouping | 19 |
| Contrast and deixis | 17 |
| Thematic highlighting | 15 |
| Quantity and scope of reference | 14 |
| Degree | 11 |
| Quality and special 'salience' | 9 |
| Disfluency and repair | 3 |
| Other | 6 |

single underlying force that may cover all its functions. Not even a sweeping reference to 'emotive highlighting' is fully adequate: this may be applicable to the intensifying and emphatic uses, but does not fit the grouping and rhythmic functions equally well and is, in any case, too general to be really helpful (which does not exclude it as a major motivating factor).

The interplay between the booster and the other pitch-prominent features, the onset and the nucleus, has only been touched on here, but it is obvious that these features may take over some of the roles played by the booster (eg in contrastive cases) and that, conversely, the booster can be said to act as a 'secondary' onset (eg in a subordinate clause following the onset) or nucleus when these features are already 'engaged' for other tasks in the tone unit. In other words, there is an interesting trade-off relationship between the pitch-prominent features in such a way that, in tone units of some length, the booster tends to be used for various 'subsidiary' functions, whether it be to express the speaker's attitude to some aspect of the utterance or to draw attention to elements carrying important information.

The latter use of the booster is most clearly revealed in its 'grouping' function, ie when it serves to indicate the beginning of a complex constituent. This function can in fact be generalized a bit further. Complex constituents almost invariably coincide with the focal element of the tone unit, and the booster can consequently be said to signal the beginning of the focus (as suggested by Cruttenden 1986:88). If we simplify a little and regard normal end-focus as beginning roughly with the first open-class word in the tone unit (provided it does not convey given information), it appears that about 75% of the examples can in fact be explained in this way. The remaining examples, which have a booster outside or inside the focus (chiefly on a contrastive thematic element or on an emphatic (asserted) transitional verb), can then be regarded as having 'marked' prominence.

However, this explanation leaves several other problems unexplained, for example why most tone units do not have a booster at all in the head of the intonation contour, while some may have as many as three, or why the booster tends to be more attracted to certain words or word classes than to others. Moreover, although textual differences have not been discussed here, the booster varies greatly in frequency from one text to another (from 127 instances in text S.1.9 to 256 instances in text S.1.6), a fact which suggests considerable situational and individual variation. (There is nothing in the material to suggest any sex-related differences, but this possibility cannot be ruled out entirely.) The booster is, it must be emphasized, a speaker-selected feature and, despite the patterns revealed here, the least predictable of the prosodic features (cf Altenberg 1987a:142). It is obvious that a great deal of further research is needed to clarify its functions in speech and its interaction with the other pitch-prominent features. What I hope to have shown here are some possible lines of inquiry and, above all, the possibilities that a collection of recorded and transcribed speech like the London-Lund Corpus offers for such research.

## Notes

1 This chapter is a revised version of a paper (Altenberg 1987c) read at the Third Nordic Conference for English Studies, Hässelby Slott, 25-27 September 1986.

2 The difference coefficient was calculated by means of the following formula (adapted from Hofland & Johansson 1982:14):

$$\frac{\text{recorded frequency} - \text{expected frequency}}{\text{recorded frequency} + \text{expected frequency}}$$

The distribution of the word categories was derived from a word-class tagged sample (ten texts totalling c 50 000 words) from the London-Lund Corpus.

# 8

# Pauses in monologue and dialogue

*Anna-Brita Stenström*

## 8.1 Introduction

When a written paragraph is read aloud, the occurrence of silent pauses is mainly influenced by graphic arrangements and the syntactic structure of the text so that, in the 'ideal delivery' of the paragraph, pauses tend to fall at natural constituent breaks (cf eg Brown 1977:91 and Clark & Clark 1977:261). In writing, the constituent structure is regularly indicated by punctuation, so that a written paragraph consists of sentences which are separated by a major punctuation mark (a period, exclamation mark, or question mark); within sentences, clauses are often separated by a comma.

Spontaneous speech is different from reading aloud. We do not always speak in full sentences, and pauses in spontaneous speech are related to features of the speech process, to the searching for words and the planning of utterances. But even though pauses in spontaneous speech signal hesitation and reflect the speaker's emotions and attitudes to a much greater extent than pauses in reading, they are bound to be affected by the syntactic structure of the utterance. Moreover, pauses in reading and pauses in speech have different realizations: in reading pauses are mainly silent, but in spontaneous speech they are silent or 'filled' (also called 'voiced'), eg *ə:m*. In spontaneous speech we also find other items that cooccur with, or substitute for, silent pauses, so-called 'verbal fillers' like *well*.

The aim of this study is threefold: first, to examine the distribution and functions of different pause types, including not only pauses 'proper', ie silent pauses (SPs), but also filled (voiced) pauses (FPs), and verbal fillers (VFs); second, to examine the relation of pause types to prosodic, linguistic and pragmatic factors; third, to investigate the extent to which pauses can be predicted.

## 8.2 Background of this study

A survey of the rich literature on pauses shows that most research has concentrated on the distribution and functions of SPs in non-spontaneous speech and, especially in the early works, on SPs in relation to sentence structure. A distinction has been made between 'juncture pauses' (ie linguistic or conventional) and 'hesitation pauses'. Lounsbury (1954), for instance, defined juncture pauses as brief (100 msec or less), falling between major constituents, and listener-oriented; and hesitation pauses as longer (up to 3 sec), occurring at 'points of lowest transition probability', and marking the beginning or end of speaker units. Boomer (1965:157), who noticed that half of the junctures in his material were followed by pauses that were 'significantly longer than hesitation pauses', questioned whether pause duration can be taken as a criterion of function. Barik (1968) observed that pauses between major constituents may be quite long and suggested that they constitute a combination of juncture and hesitation pauses.

FPs have consistently been regarded as hesitations, and even signs of stress and anxiety (cf Lallgee & Cook 1969). An additional function attributed to FPs is that of turnholder (eg Stenström 1984a). With regard to location, Maclay & Osgood (1959) observed that, in their data, FPs occurred more often before content words than before function words, whereas Cook (1971) presents data showing that FPs may occur just as often before function words. The results of Blankenship & Kay, on the other hand, indicated that hesitation pauses realized by /ah/ tended to occur before 'structural units' rather than before lexical choices (1964:369).

Except for distributional differences between silent and filled hesitation pauses, it has been suggested that there are large individual differences in FP rate, for instance due to the pressure of an audience (cf eg Cook 1971) or in a situation where the speakers cannot see each other (Kasl & Mahl 1965).

The question of encoding units seen as the effect of the relationship between the location and the function of pauses is a crucial issue in pause research. In his critical review of studies in pausology in the 50s and 60s, Rochester states that it has not been made clear 'whether pauses function in terms of words, phrases, intonation units, major grammatical constituents in the surface or deep structure, or some other aspects of utterances not described by linguistic categories' (1973:54). He speaks in favour of a multilevelled model 'in which content and theme decisions are made initially while later structural and lexical decisions proceed symmetrically' (1973:77). Similar approaches, although not as straightforwardly expressed, are found for instance in Boomer (1970) in terms of proximal and distal relationships and in Clark & Clark (1977:262ff), who suggest that speakers plan the skeleton of a sentence before its constituents. Beattie (1983:54) maintains that the main encoding units are suprasentential in scope and that speech is planned in higher-order units.

Henderson (1974) raises the question whether encoding units in speech are identified by time patterns, dismissing the thought that they should be the result of random processes. And Butterworth (1975), referring to Henderson et al (1966) and Goldman-Eisler (1968), brings up the cyclical aspect manifested in a hesitant phase directly related to the amount of phonation in the succeeding fluent phase (Butterworth 1975:76). The most detailed description of temporal patterns is provided by Beattie, who found that the mean duration of a temporal cycle consisting of a hesitant phase followed by a fluent phase was 21.88 seconds (1983:51). In the more recent literature, the emphasis is on the importance of silent (and filled) pauses for the identification of 'information units' larger than the clause (cf eg Brotherton 1979, Beattie 1983, and Chafe 1987).

Since the aim of the TESS project was to contribute to the production of more natural-sounding synthetic speech, where pauses are bound to play a crucial role, our aim was to work out a set of rules that automatically and adequately assign pauses to the synthetic-speech output of a written text. The importance of pauses for speech synthesis is stressed in Gårding (1967) who compared a sequence of recorded spontaneous speech with the same text read aloud by the original speaker and noticed that the main differences were found in variations in tempo and pausing. Predictive rules have in fact been proposed for written language read aloud. Grosjean (1980) suggests the possibility of predicting the occurrence and duration of linguistic pauses on the basis of a model that assigns to each word boundary a predicted share of the total pause duration of a sentence based on its structural complexity. A similar predictor model intended for the automatic synthesis from ordinary English text was proposed by Coker, Umeda & Browman (1973:403), who used a grammatical-category transition matrix for assigning a numerical pause potential to every word-pair boundary.

The predictability of pauses in impromptu speech is of course limited. Sigurd (1984) reports on preliminary experiments of this type for Swedish. His computer model of spontaneous speech production divides an utterance into chunks of 2-3 seconds' length, which is not governed by syntactic structure, but avoids chunking between close constituents. A pause is introduced (and a chunk finished) at the end of a sentence and may also be introduced when the system is looking for a word or planning a new sentence. The model allows simulation of spontaneous speech with different chunk lengths, filled pauses as turnholders, and speech errors.

O'Connell & Kowal (1983) conclude their extensive review of research in pausology, which covers studies as far back as the first half of the century, by asking for research focusing on 'naturalistic situations, specifically in dialogue and multilogue' instead of 'studies limited almost entirely to oral reading and speech production in contrived, artificial situations' (1983:274). Moreover, they emphasize that the majority of previous studies have concentrated on SPs without considering FPs. And their review indicates that pauses realized by VFs have scarcely even been touched upon.

Access to a large corpus of spontaneous speech should provide an excellent opportunity to describe not only the distributions and functions of SPs and FPs but also VFs, ie the mixed group of lexical items which serve as fillers of information gaps in a way similar to SPs and FPs.

## 8.3 Material and method

The material used for this study consists of ten texts from LLC (version LLC:o, see pp 19ff): one monologue (with one speaker: S.12.6) and nine dialogues (including two or more speakers: S.1.1, S.1.2, S.1.4, S.1.5, S.1.8, S.1.10, S.1.11, S.2.6, S.4.1), totalling approximately 50 000 words. A large variety of topics are discussed in the texts, ranging from academic subjects to everyday matters.

Pauses and verbal fillers are defined as follows.

**Silent pauses (SPs)**

UNIT, indicated by a dash (-) in the transcription, is the interval of an individual's rhythm cycle from one prominent syllable to the next.

BRIEF, indicated by a dot (.), is a silence perceivably shorter than unit.

LONG, indicated by two or three dashes (--, ---) or by two dashes followed by a dot (--.) are from two to three times as long as a unit pause.

### Filled pauses (FPs)

UNIT, indicated by [ə:(m)], is equivalent to a unit silent pause.
BRIEF, indicated by [ə(m)], is equivalent to a brief silent pause.

The definitions of silent and filled pauses correspond to those of Crystal (1969:166ff). Pauses in his system are of 'relative' length and depend for their definition on the overall speed of utterance for an individual. This means that the absolute duration will vary from speaker to speaker. Measurements by means of a pausemeter (described in Jönsson et al 1982 and Sigurd 1983) confirmed that, although there was some overlap, there was still a clear difference in duration not only between various pause types within a particular text but also between the same type of pause in different texts. To simplify, pauses indicated by two dashes and a dot (--•) have been included in the category of long pauses. Combinations of filled and silent pauses will be referred to as 'complex'.

### Verbal fillers (VFs)

This category includes a number of speech-specific items which are not easily accounted for as syntactic elements. In an earlier version of our tagging system they were given a code beginning with 'D' (for 'discourse') but were later referred to by AQ-tags (see p 101):

| | |
|---|---|
| HEDGES | *kind of, sort of*: he's kind of sweet |
| SOFTENERS | *I mean, you know, you see*: we can't - you know - just go away |
| STALLERS | *well*: [ə:m] well - that's what I mean |
| INITIATORS | *anyway, now* (cf 'frames' p 140), have been included here when they cooccur with pauses, although they are not fillers proper. |

The study falls into four main parts, ranging from a general survey of pauses and fillers in ten conversations to a detailed analysis of silent pauses in relation to syntactic constituents in one monologue. The last section deals with pauses and fillers as discourse phenomena.

*Part 1* (Sections 8.4-8) examines the overall distribution of pauses and of pauses in combination with VFs in ten texts.

*Part 2* (Section 8.9) accounts for the location of pauses and VFs at turn shifts and between and within tone units in ten texts.

*Part 3* (Section 8.10) focuses on the distribution of SPs and FPs in the syntactic structure of one text (the monologue).

*Part 4* (Section 8.11) discusses pauses and verbal fillers as discourse items in one-, two-, and multi-party talk.

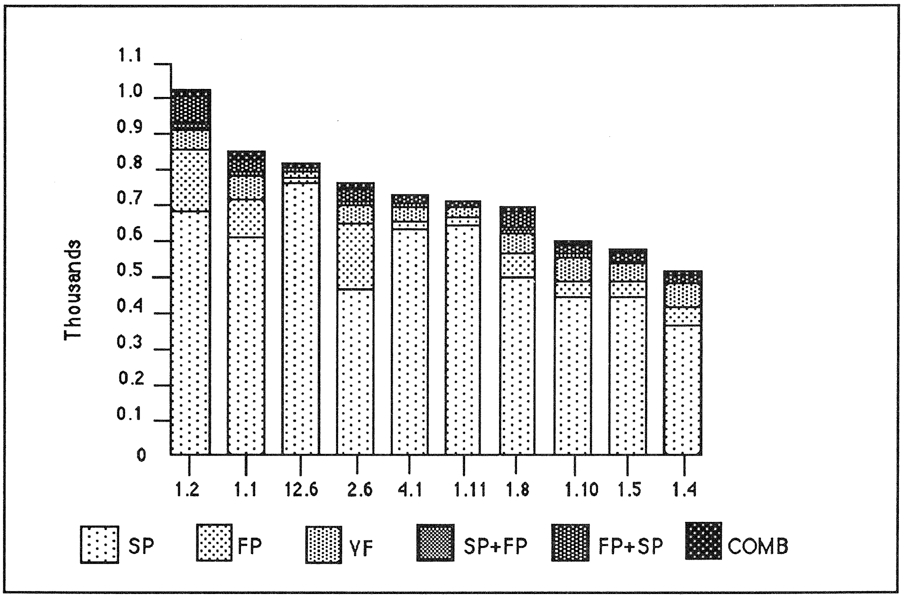
Since the study abounds in statistics, each section will be terminated by a summary of the main tendencies.

## 8.4 Frequency of pauses and verbal fillers

The density of SPs, FPs and VFs varied a great deal between the texts, both as regards the total number of pauses and fillers and the relative frequency of the different pause categories. Figure 8:1 shows that SPs dominated, that VFs and combinations of SPs and FPs were relatively rare, but that VFs + P were more frequent than SP + FP combinations. The differences in frequency are particularly obvious in text S.12.6, a prepared monologue as opposed to the rest of the texts which are all spontaneous dialogues (the texts are described in Chapter 1, Appendix 1). But the difference in distribution between some of the dialogues is also notable. Why, for instance, does text S.1.2 contain so many SPs and text S.1.4 so few? And why is the difference in occurrence between SPs and FPs so small in text 2.6? One reason is of course individual speaker habits: some speakers pause more than others; and some use FPs while others do not. Another reason is the speech situation: the speakers may be on more or less intimate terms and the conversations may range from serious discussions to informal chats. In each of the three subtexts making up text S.1.2, for instance, two male academics discuss strictly academic matters; in text S.1.4 a couple of colleagues chat about less serious matters while choosing pictures for the department; in text S.2.6, with more FPs than in the other dialogues, the four speakers seem to insert an FP in strategic places in their attempts to get the turn. The overall distribution is shown in Table 8:1. The total number of pauses and verbal fillers per text varied from 1036 in text S.1.2 to 525 in text S.1.4. Note that text S.12.6 contained the third largest number of pauses (and verbal fillers).

The proportion of SPs was very high (76%) compared to that of FPs (10% including glottal FPs) and cooccurrences of SP and FP (5%), but there were considerable variations within the texts. The largest proportion of SPs occurred in text S.12.6 (96%) the smallest in text 2.6 (60%). The distribution in texts S.1.4 and S.12.6 is of special interest, since much of the remaining discussion concentrates on pauses in these two texts.

**Figure 8:1**. Distribution of SPs, FPs and VFs from 10 texts.




The fact that FPs were far less frequent than SPs (cf Crystal 1969:167), may be an indication of their basically different functions: SPs serve typically as juncture pauses, while FPs primarily indicate hesitation.

## 8.5 The ratio of pauses to number of words and tone units

The proportion of SPs, FPs, and VFs in relation to the number of words and tone units is shown in Table 8:2 (where no distinction has been made between types of pause and between categories of filler). The large variation in words/pause ratio and tone units/pause ratio is striking: words per SP range from 13.5 to 6.3 and tone units per SP from 3.3 to 1.6; words per FP range from 416.7 to 27.3 and tone units per FP from 102.3 to 7.1; words per VF range from 238.1 to 44.6 and tone units per VF from 81.8 to 13.7.

How do we account for this variation? Let us consider text S.12.6 with its extreme figures. As shown in Figure 8:1, this text contains more SPs than the dialogues. Consequently the words/SP and tone units/SP ratios are low. By contrast, it has fewer FPs and VFs than the dialogues, and therefore the ratios for FP and VF are high. This means that there is a large proportion of silence in text S.12.6, which in turn results in comparatively slow speech.

**Table 8:1.** SPs, FPs, and VFs in ten texts.

| P-TYPE | 1.2 | 1.1 | 12.6 | 2.6 | 4.1 | 1.11 | 1.8 | 1.10 | 1.5 | 1.4 | TOTAL |
|---|---|---|---|---|---|---|---|---|---|---|---|
| SP | 692 | 614 | 789 | 466 | 643 | 649 | 500 | 447 | 444 | 370 | 5 614 |
| % | 67 | 71 | 96 | 60 | 87 | 89 | 69 | 73 | 76 | 70 | 76 |
| FP | 162 | 106 | 12 | 174 | 11 | 23 | 68 | 35 | 36 | 56 | 683 |
| % | 16 | 12 | 1 | 22 | 1 | 3 | 9 | 6 | 6 | 11 | 9 |
| GLOTT FP | 16 | 3 | 0 | 9 | 2 | 2 | 3 | 10 | 3 | 4 | 52 |
| % | 2 | 0.3 | | 1 | 0.3 | 0.3 | 0.4 | 2 | 0.4 | 1 | 1 |
| VF alone | 56 | 74 | 13 | 59 | 56 | 44 | 73 | 70 | 70 | 71 | 586 |
| % | 5 | 9 | 2 | 8 | 8 | 6 | 10 | 11 | 12 | 13 | 8 |
| SP+FP | 25 | 12 | 3 | 10 | 4 | 2 | 14 | 5 | 2 | 0 | 77 |
| % | 2 | 1 | 0.4 | 1 | 1 | 0.3 | 2 | 1 | 0.3 | | 1 |
| FP+SP | 71 | 38 | 4 | 55 | 17 | 10 | 57 | 40 | 21 | 20 | 333 |
| % | 7 | 4 | 0.5 | 7 | 2 | 1 | 8 | 7 | 4 | 4 | 4 |
| COMBINATIONS | 14 | 13 | 0 | 9 | 2 | 2 | 6 | 3 | 6 | 4 | 59 |
| % | 1 | 1 | | 1 | 0.3 | 0.3 | 1 | 0.5 | 1 | 1 | 1 |
| TOTAL | 1036 | 860 | 821 | 782 | 735 | 732 | 721 | 610 | 582 | 525 | 7404 |
| % | 14 | 12 | 11 | 10 | 10 | 10 | 10 | 8 | 8 | 7 | |

If the average number of words and tone units per SP, FP, and VF is calculated, we get the following result:

| | | | |
|---|---|---|---|
| Words per SP: | 9.3 | Tone units per SP: | 2.4 |
| Words per FP: | 149.6 | Tone units per FP: | 36.3 |
| Words per VF: | 71.4 | Tone units per VF: | 24.3 |

However, the average number of words and tone units per pause type tells us nothing about the exact location of the pauses. These figures do not show, for instance, that approximately every second tone unit is preceded by an SP, or that speakers' planning embraces 2.4 tone units. (This in turn presupposes that an interpausal unit is coextensive with a planning unit, which is not necessarily true.) What we might say is that the speaker's 'performance units' (cf Section 8.9.6) have a certain average length. But exactly what is to be found in the units is still unclear.

In this connection, it may be mentioned that Little (1963:50) found that 'vocalized pauses' realized by *well* and *uh* in sophisticated speech occurred at a rate of one per 25 words; a rough estimation of the frequency of *well* + FP in the present data indicates one occurrence per 16 words. However, since we are not trying to produce rules for 'natural' pauses in impromptu speech but for written English read aloud, our main concern is the frequency and location of silent pauses, not filled pauses or verbal fillers, which are typical of spontaneous speech. Notice the low density of FP and VF in text S.12.6, which comes closest to reading in the corpus (Table 8:1).

However, brief SPs (represented by a single dot in the transcription) make a special case in sometimes being so brief as to be difficult to identify unambiguously and will not have a high priority for pause prediction. Therefore, I made a special study of the distribution of SPs that were longer than brief in relation to the number of words and tone units. The average number of words per SP longer than brief was 22.7 and the average number of tone units per SP longer than brief was 5. Note that the comparatively low number of words and tone units per SP in text S.12.6 (Table 8:2) is the effect of the large proportion of brief SPs. We will now look at the distribution of SPs, FPs, and VFs in greater detail.

**Table 8:2**. Rates of SP, FP, and VF in relation to words and TUs.

| TEXT | WORDS PER SP | TUs PER SP | WORDS PER FP | TUs PER FP | WORDS PER VF | TUs PER VF |
|---|---|---|---|---|---|---|
| S.1.4 | 13.5 | 3.3 | 83.3 | 20.5 | 48.5 | 14.6 |
| S.1.5 | 11.3 | 2.9 | 128.2 | 33.6 | 45.9 | 15.9 |
| S.1.10 | 11.2 | 2.9 | 111.1 | 29.0 | 44.6 | 16.1 |
| S.2.6 | 10.7 | 2.8 | 27.3 | 7.1 | 59.5 | 19.5 |
| S.1.8 | 10.0 | 2.3 | 70.4 | 16.0 | 45.9 | 13.7 |
| S.1.1 | 8.1 | 1.9 | 45.9 | 11.1 | 48.5 | 14.6 |
| S.1.11 | 7.7 | 1.9 | 200.0 | 50.6 | 75.8 | 25.3 |
| S.1.2 | 7.2 | 2.1 | 28.1 | 8.2 | 50.0 | 23.9 |
| S.4.1 | 7.8 | 1.7 | 384.6 | 84.5 | 57.5 | 17.2 |
| S.12.6 | 6.3 | 1.6 | 416.7 | 102.3 | 238.1 | 81.9 |

## 8.6 Silent and filled pauses

More than half the total number of SPs were brief. And the longer the pause the less frequent it was, as is reflected in the decreasing numbers in Table 8:3, with the overall frequencies of 58% brief, 27% unit, and 15% long for silent pauses. There are large variations in pause length between individual texts. Compare for instance the monologue (S.12.6) and text S.4.1 (a dialogue), both with 26% unit SPs but with very different numbers of long SPs (9% and 27%, respectively).

Considering that brief SPs were much more frequent than unit SPs, it is interesting to note the reverse tendency for FPs. Table 8:4 shows that 48% were unit and 41% brief FPs. This, however, is a consequence of the large number of unit FPs in two of the texts (S.1.2 and, especially, S.2.6). The fact that, in some texts, brief FPs were more common than unit FPs indicates that the appearance of one or the other pause type is very much a matter of speaker habit and speech situation. In combinations of silent and filled pauses, the general tendency was for FP to precede SP. Cases with SP preceding FP were comparatively few in most of the texts. Both types are illustrated in (1):

(1) is is [ə] • ‖Mallet has pro▵duced a [ə:m] a {RE△VÌSED} CONSTI▵TÙTION■ ‖FÒR■ [ði] ‖School of YÍDDISH■ • in ‖which • [ə:m] the main △PÒINT■ of my (...) (S.1.2:22-23)

FP generally preceded SP at the beginning of the utterance, but not necessarily in initial position as in (1), and SP usually preceded FP within the utterance:

in ‖which • [ə:m] the main △PÒINT■

In the first case FP is used as a turntaker, in the second as a turnholder (the location of pauses in relation to turntaking will be discussed in detail in Sections 8.3 and 8.5). The tendency for SP to follow FP is a natural consequence of the composition of my corpus; nine of the ten texts were dialogues with FPs generally appearing in turn-initial position. The monologue contained only four instances of the FP + SP combination.

Although combinations of SP and FP can be quite complex, they were, on the whole, rather infrequent in this material. The most common types of combination were SP+FP+SP and FP+SP+FP. The dialogue was alone in having no such combinations.

**Table 8:3**. Distribution of SPs.

| SP-<br>LENGTH | TEXTS<br>12.6 | 1.2 | 1.11 | 4.1 | 1.1 | 1.8 | 2.6 | 1.10 | 1.5 | 1.4 | TOTAL |
|---|---|---|---|---|---|---|---|---|---|---|---|
| BRIEF | 516 | 464 | 312 | 301 | 390 | 280 | 290 | 227 | 237 | 235 | 3252 |
| % | 65 | 67 | 48 | 47 | 63 | 56 | 62 | 51 | 53 | 64 | 58 |
| UNIT | 203 | 169 | 187 | 169 | 139 | 143 | 136 | 130 | 141 | 105 | 1522 |
| % | 26 | 24 | 29 | 26 | 23 | 29 | 29 | 29 | 32 | 28 | 27 |
| LONG | 70 | 59 | 150 | 173 | 85 | 77 | 40 | 90 | 66 | 30 | 840 |
| % | 9 | 9 | 23 | 27 | 14 | 15 | 9 | 20 | 15 | 8 | 15 |
| TOTAL | 789 | 692 | 649 | 643 | 614 | 500 | 466 | 447 | 444 | 370 | 5614 |
| % OF ALL TEXTS | 14 | 12 | 12 | 11 | 11 | 9 | 8 | 8 | 8 | 7 | 100 |

**Table 8:4**. Distribution of FPs.

| P-TYPE | TEXTS<br>2.6 | 1.2 | 1.1 | 1.8 | 1.4 | 1.10 | 1.5 | 1.11 | 4.1 | 12.6 | TOTAL | % |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| BRIEF<br>[ə(m)] | 33 | 87 | 54 | 32 | 41 | 17 | 9 | 12 | 10 | 6 | 301 | 41 |
| UNIT<br>[ə:(m)] | 132 | 69 | 50 | 33 | 12 | 17 | 25 | 10 | 1 | 6 | 355 | 48 |
| FP COMB<br>[ə(m)] + | 6 | 5 | 2 | 2 | 1 | 1 | 2 | | | | 19 | 3 |
| [ə:(m)] + | 3 | 1 | | 1 | 2 | | | 1 | | | 8 | |
| GLOTTAL FP | 9 | 16 | 3 | 3 | 4 | 10 | 3 | 2 | 2 | 0 | 52 | 7 |
| TOTAL | 183 | 178 | 109 | 71 | 60 | 45 | 39 | 25 | 13 | 12 | 735 | |

## 8.7 Verbal fillers

In spontaneous conversation, VFs often cluster with SPs and FPs as illustrated in (2):

(2) and ||ALL this was DONE [ə:]■ -- ||by -- ▻kind of ▻letting - [ə:] ---• ||{WELL}
||REALLY by 'just [ə:] --'sort of [ə]■ - ||starting from △NOTHING■ (S.2.3:115-117)

The example may seem exaggerated but is by no means unique. Goldman-Eisler (1968) found that the proportion of silence in relation to the total speaking time ranged between 13% and 67% in her impromptu talk data. This example shows that long silences do not exclude but rather invite other hesitation phenomena. The distribution of VFs and pauses in the ten texts is presented in Table 8:5. Generally speaking, VFs occurred more often alone than accompanied by a pause.

The monologue had a different distribution of VFs than the dialogues with a very low total number of VFs (and a pause). Only the initiator *now* and the staller *well* were fairly well represented, with the latter used in utterance-initial position like the initiators *anyway* and *now*, eg to resume or shift topics:

(3) -- ||well Stoke 'Court as you △KNOW■ (S.12.6:699)

**Table 8:5.** VFs with and without a pause (P) in ten texts (+ P means any order of VF and P).

| VF-TYPE | 1.10 | 1.8 | 1.5 | 1.4 | 1.1 | 1.2 | 4.1 | 2.6 | 1.11 | 12.6 | TOTAL | % |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| SOFTENER | | | | | | | | | | | | |
| ALONE | 44 | 41 | 26 | 25 | 35 | 22 | 18 | 26 | 11 | 1 | 249 | 28 |
| + P | 29 | 30 | 17 | 22 | 15 | 25 | 16 | 10 | 10 | - | 174 | 19 |
| WELL | | | | | | | | | | | | |
| ALONE | 15 | 28 | 26 | 42 | 35 | 29 | 31 | 29 | 26 | 8 | 269 | 30 |
| + P | 8 | 4 | 14 | 10 | 8 | 13 | 10 | 13 | 9 | 3 | 92 | 10 |
| HEDGE | | | | | | | | | | | | |
| ALONE | 10 | 4 | 16 | 4 | 1 | 3 | 5 | 2 | 5 | - | 50 | 6 |
| + P | 5 | 2 | 7 | - | 3 | 2 | 4 | 1 | 2 | - | 26 | 3 |
| INITIATOR | | | | | | | | | | | | |
| ALONE | 1 | - | 2 | - | 3 | 2 | 2 | 2 | 2 | 4 | 18 | 2 |
| + P | - | - | 1 | - | 3 | 4 | 1 | 1 | 1 | 5 | 16 | 2 |
| TOTAL | 112 | 109 | 109 | 103 | 103 | 100 | 87 | 84 | 66 | 21 | 894 | |

It was SPs (rather than FPs and complex pauses) that occurred with VFs, probably because VFs can fill the gaps in the same way as FPs and complex pauses. Softeners, which were more often found turn-medially than in other positions, were accompanied by brief SPs in most cases. This points to their role as emphasizers rather than hesitators. The hedge *sort of* was accompanied by a unit SP as often as a brief SP. Initiators cooccurred with unit SPs more often than brief SPs and also fairly often with unit FPs, maybe a reflection of their frequent position in the hesitation area at the beginning of a turn. Surprisingly, *well* was more often accompanied by a brief SP than a unit SP, which seems to indicate that it was used as a response-initiator signalling 'insufficiency' more often than as a staller signalling hesitation. On the other hand, *well* was found with long and complex pauses more often than other VF categories, which points to its use in hesitation areas.

In most of the cases the pause preceded VF. The exceptions were *sort of*, which usually came immediately before a pause, and *you see* which, unlike *you know*, more often preceded the pause (for the difference in functions, see Svartvik & Stenström 1985 and Erman 1987).

## 8.8 Summing up

- We can note that the total occurrence of pauses per individual text ranged between 14% and 7% (Table 8:1).
- In the individual texts, SPs ranged between 96% and 60% of the total number; FPs (including glottal FPs) from 23% to 1%; VFs from 13% to 2% (Table 8:1).
- 76% of the pauses in the ten texts consisted of SPs alone; 10% consisted of FPs alone (including glottal FPs); 8% consisted of VFs alone (Table 8:1).
- 58% of all SPs were brief and 42% unit or longer; the figures are almost exactly the opposite for FPs (Table 8:3).
- FPs were less frequent than SPs and generally also than VFs in relation to words and tone units; exceptions were texts S.2.6, S.1.1 and S.1.2 (Table 8:2).
- SPs served typically as juncture pauses; FPs primarily indicated hesitation.
- FP tended to precede SP at the beginning of an utterance; SP preceded FP within the utterance.
- Brief (rather than unit or long) SPs were followed by FP; unit (rather than brief) FPs were followed by SP.
- One in three VFs was accompanied by SP or FP or by a combination of both and more often by SP (generally brief) than FP.

- The pause (generally SP) preceded VF in most cases; exceptions were *sort of* and *you see*.
- *Anyway* and *now* attracted the highest and *well* the lowest proportion of pauses in relation to their total number, 89% and 34% respectively (Table 8:5).

## 8.9 The location of pauses and verbal fillers

The location of pauses and verbal fillers has been examined from two main points of view:

| | |
|---|---|
| Turntaking: | To what extent do speakers pause at turn shifts and what types of pause do they use then? |
| Turn organization: | Once the speaker has taken the turn, how does he organize his speech, judging by the way he pauses? |

Since the distribution of pauses involving VFs is markedly different in monologue and dialogue structure, each subsection begins with a general survey of the location in the ten texts before narrowing down to a comparison of occurrences in the monologue (text S.12.6) and one of the dialogues (text S.1.4).

For the comparative study of these two texts it will be useful first to take a look at their general characteristics. Text S.12.6 was selected since it was found to be the monologue in the corpus that came closest to reading, and text S.1.4 was chosen for contrast (as one of many dialogues). The two texts differ especially in the following respects:

- The monologue contains no interruptions from a second party other than in the form of laughter; it is a basic characteristic of the dialogue to contain turntaking.
- The monologue is unscripted but prepared, and maybe partly read; the dialogue is spontaneous.
- Both texts contain a number of one-word tone units; in the monologue such tone units contain words like *and*, *but*, *so*, and *now* which link clauses together and serve to carry the talk forward; in the dialogue the majority of the one-word tone units consist of feedback signals, such as *yes*, *no*, and *m*.

- The speaker in the monologue is exceptionally slow; both parties in the dialogue are fast speakers and fast turntakers, which partly explains the differences in tone unit segmentation and pause distribution.

The amount of speech (reflected in the number of tone units produced in a certain period of time) was more than twice as large in the dialogue, with 1227 tone units in 22 min 20 sec, as in the monologue, with 1211 tone units in 47 min 50 sec (cf Altenberg 1987a:22). It is possible, however, that the low speech rate in the monologue is not the effect of high pause frequency alone but of pause rate in combination with 'short-unit segmentation' involving shorter than brief 'pauses' at tone unit boundaries (cf Crystal 1969:171); fewer words per tone unit were produced in the monologue than in the dialogue (not including feedback signals and instances of simultaneous speech which cause a skewed distribution). This suggests that segmentation into shorter tone units *per se* had a speed-reducing effect, a matter that I have not looked into here.

Examples (4) and (5) illustrate the overall difference in tone unit segmentation in the two texts, with a pause separating each tone unit in (4) but not in (5):

(4) the ||Hanbury FÀMILY■ – ||LÌVED THÉRE■ • with ||five CHÌLDREN■ – and ||she ΔÀCTUALLY■ • ||finished her ΔEDUΔCÀTION■ • in ||what is KNǑWN■ ||as the servants' ▷HÁLL■ (S.12.6:412-418)

(5) **||I was** [ə] ||I was [ə:m] • re||duced just to 'putting "Δone of ΔÈACH■ but [ði it] it "||does 'spell it ÒUT■ if you've got *"||two* PÀSSAGES■ (S.1.4:357-359)

### 8.9.1 Pauses and verbal fillers at turn boundaries and within turns

Since pauses and fillers at turn boundaries can only be observed in dialogues, no comparison with text S.12.6 has been made in this section.

VFs accompanied by a pause (before or after) occurred at turn boundaries and within turns in the nine dialogues as shown in Table 8:6. Softeners in combination with a pause were most often found within the turn. This result was expected for *I mean*, which often serves to introduce an afterthought or an explanatory remark, but not for *you see* and *you know*, which usually occur in turn-final position inviting feedback. The turn-medial position of *you see* and *you know* can therefore be taken to indicate either that the listener gave silent instead of oral feedback, or that the feedback consisted of minimal

**Table 8:6.** Position of VF+P in the turn in nine dialogues ('medial' is anywhere between the first and the last word).

| VF-TYPE | INITIAL | MEDIAL | FINAL | TOTAL |
|---|---|---|---|---|
| SOFTENER+P | | | | |
| *you know* | 9 | 66 | 22 | 97 |
| *you see* | 5 | 22 | 7 | 34 |
| *I mean* | 6 | 37 | | 43 |
| WELL+P | | | | |
| *well* | 43 | 44 | 2 | 89 |
| HEDGE+P | | | | |
| *sort of* | | 26 | | 26 |
| INITIATOR+P | | | | |
| *anyway* | 2 | 1 | | 3 |
| *now* | 7 | 1 | | 8 |
| TOTAL | 72 | 197 | 31 | 300 |

responses like *mhm*, *yes*, *right*, which are no proper turns, or that the feedback came at the same time as *you see* and *you know*.

Surprisingly enough, the figures indicate an equal distribution of *well* + P within and at the beginning of a turn. One would have expected *well* to be much more frequent in turn-initial position, considering its typical function as a response-initiator (for a different distribution, see Little 1963:49 and Svartvik 1980a). Moreover, 'within' in Table 8:6 includes *well* + P in second position, eg *[ə:m] well*, *yes well*, etc. The hedge *sort of* + P was not found at turn-boundaries in this data. The initiators *anyway* + P and *now* + P occurred as topic/aspect shifters both turn-initially and within the turn.

**Table 8:7.** SPs and FPs at turn boundaries in Text S.1.4.

| P-TYPE | BRIEF | UNIT | LONG | COMBINATIONS | TOTAL |
|---|---|---|---|---|---|
| SILENT | 35 | 18 | 5 | | 58 |
| FILLED | 3 | 2 | | | 5 |
| COMPLEX | | | | 5 | 5 |
| TOTAL | 38 | 20 | 5 | 5 | 68 |

The distribution of SPs and FPs at turn boundaries in text S.1.4 is shown in Table 8:7. The fact that there were such frequent interruptions, with speakers not waiting for their turn, explains to some extent the few pauses in connection with speaker shifts. This results in what Oreström calls 'unsmooth' speaker shifts (1983:138-71), as in (6):

(6) B: well it's ||sort of ΔTÒO • ||YÈS■ *it's*

A: *I* ||MĒAN■ it ||would ▻be a 'bit 'out of ΔPLÀCE *SÓMEHOW*■ (S.1.4:479-482)

Here A starts his turn before B has finished (**I** is simultaneous with **it's**). Oreström found that, in his data, consisting of ten face-to-face LLC dialogues, only 14.8% of the turntakings were unsmooth (1983:166).

Whether SPs between turns should be attributed to the current or next speaker is an open question. However, it might be assumed that turn-final SPs generally serve as turnyielders in ordinary conversation as opposed to, for instance, interviews, where the interviewee makes a pause before answering a tricky question. With FPs, on the other hand, it was quite obvious that they served as turn initiators:

(7) A: [ə:m] • the the ||point ÌS■ that ||this «has got [igz]» [ə] has been ▵filled up with EXΔAMINǍTION ▻papers■ *-* ||you SÉE■ *-* (S.1.2:866-868)

### 8.9.2 Pauses and verbal fillers at tone unit boundaries

Of all instances of VF + P just over one third came immediately after a tone unit boundary. The remaining two thirds were found immediately before a tone unit boundary or in post-onset position. The distribution in the ten texts is presented in Table 8:8, which however does not specify whether the pause came before or after VF. The difference in position between *you see* and *you know* + P on the one hand and *I mean* + P on the other is worth noticing. *You see* and *you know* + P typically preceded a tone unit boundary, either terminating the information contained in the tone unit:

(8) is sort of ||begging for the ΔMÒON■ -- «you ||SÉE■» (S.1.1:743-744)

or introducing the new information provided in the next tone unit:

(9) [ə:m]-- ||and [ə:m] you KNÓW■ - ||if Δthis is ▵ÀLSO CÓME■ ||from [ə] - NÍGHTINGALE • (...) (S.1.1:268-271)

**Table 8:8.** VF+P/P+VF immediately following a tone unit boundary in ten texts.

| VF-CATEGORY | AFTER TU BOUNDARY | OTHER POSITION | TOTAL |
|---|---|---|---|
| SOFTENER+P | | | |
| *you know* | 25 | 72 | 97 |
| *you see* | 1 | 33 | 34 |
| *I mean* | 25 | 18 | 43 |
| WELL+P | | | |
| *well* | 36 | 56 | 92 |
| HEDGE+P | | | |
| *sort of* | 20 | 6 | 26 |
| INITIATOR+P | | | |
| *anyway* | | 4 | 4 |
| *now* | 5 | 7 | 12 |
| TOTAL | 112 | 196 | 308 |

*I mean* + P typically occurred after a tone unit boundary (in pre-onset position), also introducing new information in the next tone unit:

(10) it's -- ▵LÀNGUAGE■ --- I mean ‖my i▵dea would (...) (S.1.1:292-293)

but also in post-onset position with the same function:

(11) [ə:m] -- «‖well I mean» • the ‖way these chaps ▵GÒ■ (S.1.1:595)

One moot point in intonation studies is whether tone units are by definition, or typically, separated by pauses. Or should the presence or absence of pauses between tone units be described as speaker/topic/situation specific? One way of avoiding the problem altogether would be to opt for 'pause-defined units' instead of 'contour-defined' tone units, as suggested by Brown, Currie & Kenworthy 1980:69ff), ie chunks of speech bounded by SPs instead of tone unit boundaries. However, since LLC has been analysed in terms of tone units, this is not the place to go into this general question. Yet the corpus can be used for studying the relation between pauses and tone unit boundaries.

To get a rough idea of the distribution of SPs and FPs at tone unit boundaries I examined the first, middle, and last one hundred tone units in S.12.6 and S.1.4, and found that only slightly more than half of the tone units in the former cooccurred with a pause and no more than one in four in the

**Table 8:9.** Tone unit boundaries with (+P) and without (-P) a following pause.

| TEXT S.12.6 | | TEXT S.1.4 | |
|---|---|---|---|
| TU boundary -P | TU boundary +P | TU boundary -P | TU boundary +P |
| 127 | 173 | 240 | 60 |

latter. The results appear in Table 8:9, in which figures from the first, middle and final one hundred tone units in each text have been conflated. Pauses cooccurred with tone unit boundaries almost three times as often in the monologue as in the dialogue. The most obvious reason for this is that pauses are less likely to be found in a dialogue with generally very short turns, often consisting of just a backchannel item making up a one-word tone unit, than in a monologue where the same speaker goes on speaking for nearly 50 minutes. Brief SPs dominated in both conversations while FPs and complex pauses were extremely rare. On the basis of this sample, then, the answer to the question 'Are tone units typically separated by pauses?' is that it depends on whether the talk occurs in a monologue or a dialogue.

### 8.9.3 Pauses and verbal fillers within tone units

Verbal fillers may make up separate tone units, with or without pauses. Table 8:10 shows how often VF + P made up a separate tone unit and the location within the tone unit (no distinction has been made between cases where P preceded and followed VF). VF + P constituted a separate tone unit in more than one third of the cases:

(12) *‖WÈLL [ə:]■ when I ‖FÌRST did* ARISTÓPHANES■ (S.1.4:275-276)

Only *sort of* + P did not appear in that position:

(13) ‖sort of • ▵SÈRVICE FLÁTS■ (S.1.10:1039)

Another third appeared in the middle of a tone unit:

(14) but ‖it's [əm] – I mean it's ‖got SHÁPE■ (S.1.8:658)

**Table 8:10.** Position of VF and pause within tone units.

| | |
|---|---|
| ONSET POSITION: | Indicates that the items occurred immediately before or immediately after onset without constituting a separate tone unit. |
| TU MEDIAL: | Denotes position between 'onset position' and 'TU boundary follows'. |
| TU BOUNDARY FOLLOWS: | Indicates that the items occurred immediately before a tone unit boundary. |

| VF-TYPE | ONSET POSITION | | MEDIAL IN TU | | TU BOUNDARY FOLLOWS | | SEPARATE TU | | TOTAL |
|---|---|---|---|---|---|---|---|---|---|
| SOFTENER+P | | | | | | | | | |
| *you know* | 10 | 10% | 25 | 26% | 4 | 4% | 58 | 60% | 97 |
| *you see* | 4 | 12% | 1 | 3% | 8 | 24% | 21 | 62% | 34 |
| *I mean* | 15 | 35% | 25 | 58% | 1 | 2% | 2 | 5% | 43 |
| WELL+P | | | | | | | | | |
| *well* | 24 | 26% | 36 | 39% | 1 | 1% | 31 | 34% | 92 |
| HEDGE+P | | | | | | | | | |
| *sort of* | 6 | 23% | 20 | 77% | | | | | 26 |
| INITIATOR+P | | | | | | | | | |
| *anyway* | | | | | | | 4 | 100% | 4 |
| *now* | 6 | 50% | 5 | 42% | | | 1 | 8% | 12 |
| TOTAL | 65 | 21% | 112 | 36% | 14 | 5% | 117 | 38% | 308 |

The least favoured position was immediately before a tone unit boundary, which is explained by the fact that the only VF-items that usually appear in that position are *you know* and *you see*:

(15) ||MOSTLY it 'means that■ the ||kids AREN'T • {you ||SEE■ ||}■ (S.1.10:93-96)

Looking at individual VF-items + P we notice that most instances of *you see* + P and *you know* + P constituted a separate tone unit (62% and 60%, respectively). *I mean* + P occurred more frequently in medial position, while *well* was found in onset position almost as often as in a separate tone unit. Finally, *anyway* + P unlike *now* + P occurred only in a separate tone unit:

(16) ||ANYWAY■ • shall we ||TURN to (...) (S.2.6:380-381)

(17) – • now ||poor Sir FREDERICK■ – got ||mixed UP■ (S.12.6:623-624)

The total number of pauses within the tone unit was almost the same in the two texts. SPs constituted the most common pause type in both, although SPs were relatively more frequent in the monologue. Table 8:11 presents the total distribution of SPs and FPs, and complex pauses between and within tone units. Pauses in combination with speaker shifts in text S.1.4 (68 instances, see Table 8:7) are excluded. Pauses between tone units were by far the most frequent in both texts, realized by SPs in the majority of the cases. But there is an interesting difference in proportion within the texts: twice as many pauses occurred between as within tone units in the dialogue compared with four times as many in the monologue. This points to a relatively higher degree of concurring performance and tone units in the monologue than in the dialogue. Although the monologue contained twice as many SPs as the dialogue, the difference in number *within* the tone unit was less significant. This only serves to stress that the difference in SP frequency is due to pauses occurring *between* the tone units.

When it comes to FPs the result was the opposite: the dialogue contained four times as many FPs as the monologue (75% of which occurred within the tone unit in the dialogue, as compared with 67% in the monologue). Since FPs are hesitation signals in the first place, this shows not only that the two parties in the spontaneous dialogue hesitated more often than the speaker in the preplanned monologue but also that hesitation is a very 'local' phenomenon. The fact that the total distribution of silent, filled, and complex pauses was less uneven in the dialogue than in the monologue highlights the different speech situations.

**Table 8:11.** Distribution of SPs and FPs in a monologue and a dialogue.

| PAUSE | MONO | DIA |
|---|---|---|
| WITHIN TUs | | |
| silent | 144 | 69 |
| filled | 8 | 37 |
| complex | 4 | 12 |
| TOTAL | 156 | 118 |
| BETWEEN TUs | | |
| silent | 645 | 245 |
| filled | 4 | 15 |
| complex | 3 | 8 |
| TOTAL | 652 | 268 |
| TOTAL | 808 | 386 |

### 8.9.4 Summing up

- *I mean*, *you know* and *you see*, and *sort of* in combination with a pause, were found more often in medial than in initial and final position, but not *anyway*, *now*, and *well* (Table 8:6).
- In the dialogue, less than every fourth speaker-shift had a pause, usually brief and silent (Table 8:7).
- FPs occurred in turn-initial position; whether SPs were turn-initial or turn-final could not be determined.
- VFs were less often found immediately after a tone unit boundary than anywhere else (Table 8:8).
- The agreement between tone unit boundaries and occurrence of pauses was unexpectedly low (Table 8:9).
- FPs were more frequent within TUs, SPs between tone units (Table 8:11).
- *You know*, *you see*, and *anyway* occurred more often in a separate tone unit; *I mean*, *well*, and *sort of* were more often found in medial position; *now* (as opposed to *anyway*) preferred onset position (Table 8:10).

## 8.10 Pauses as linguistic demarcators

One of the aims of this study was to suggest rules for automatic pause assignment based on the occurrence of pauses in genuine speech, notably pauses separating syntactic constituents. With this in mind I made a special study of one of the texts. Since it was preferable to use data that was as close as possible to writing, I selected text S.12.6, which is largely free from the hesitations, reformulations, and anacolutha that are typical of impromptu speech. Clearly, a far more extensive material is needed for writing reliable predictive rules for pause assignment. Therefore I shall only point to the main tendencies in this text. (For previous research in this area, see Section 8.1.)

Strings of words delimited by pauses will be referred to as 'performance units' (cf Section 8.10.6). Such strings, which may or may not run across tone unit boundaries, are free from internal pauses and roughly equivalent to what Beattie (1983) refers to as 'fluent units'. Since pauses operate not only at the syntactic level but also at the discourse level, Section 8.10 will be devoted to pauses as 'discourse markers'.

Both SPs and FPs separated syntactic constituents and pauses were classified according to whether they were found **between sentences**, ie where a full stop would be likely in writing; **between clauses**, ie where a full stop would not be likely in writing; **between the clause elements S, V, C, O, and A** (Quirk et al 1985:49); **between phrase elements**, ie words in noun phrases

(NPH), verb phrases (VPH), adjective phrases (JPH), adverb phrases (APH), prepositional phrases (PPH); **with conjunctions.**

The four-level tagging of the text helped to determine which of these levels were affected by the pause (cf Chapter 4). This is demonstrated in Figure 8:2.

(a) a unit SP separates a conjunction from a preposition that is part of a prepositional phrase functioning as an adverbial;
(b) a unit SP separates a determiner from its head in the noun phrase;
(c) a unit SP separates a verb from its object;
(d) a unit SP separates the initiator *now* from the rest of the discourse.

Note that although (a) and (d) are very similar they are not identical; *that* in (a) is part of a clause containing a fronted adverbial:

(18) the most ‖modern 'farm ▵BUILDINGS■ that – ‖in that PERIOD■ were ‖rather U▵NIQUE■ (S.12.6:594-596)

**Figure 8:2.** Four-level tagging.

| | | | | | | |
|---|---|---|---|---|---|---|
| (a) | | that | – | in | ‖that | PERIOD■ |
| | WORD-CLASS level | CD | | PA | TD | NC |
| | | | | PPH | | |
| | | S | | A | | |
| (b) | | ‖leaving | all | his | – | CARS■ |
| | | VA+G | EC | TB | | NC+2 |
| | PHRASE level | VPH | NPH | | | |
| | | V | O | | | |
| (c) | | it | would | INTEREST | – | ME |
| | | RC | VM+9 | VA+O | | RB |
| | | NPH | VPH | | | NPH |
| | CLAUSE level | S | V | | | O |
| (d) | | ‖NOW | – | ‖AFTER | the | ALLUYSONS■ |
| | | DI | | PA | TA | NP+2 |
| | | | | PPH | | |
| | | | | A | | |
| | DISCOURSE level | INIT | | | | |

But *now* could not be incorporated in the following clause:

(19) ||NOW■ - ||AFTER • the ALLUYSONS■ • [ə:] Sir ||Philip ▵WILD■ (...) ||stopped • and ▵LIVED at 'this par'ticular MANSION■ (S.12.6:733-737)

*Now* is here a self-contained item which serves to initiate a new stage in the narrative and belongs to a separate level, referred to as the 'discourse level' in our system.

When examining pauses in the syntactic hierarchy it was necessary to go across tone unit boundaries. The fact that the tone unit is regarded as the basic prosodic (and information) unit does not necessarily imply that tone units and syntactic units are related in a one-to-one fashion. 81% of all SPs in text S.12.6 occurred between tone units, but only 19% within tone units (Table 8:12). The majority of all SPs were brief, followed by unit SPs, while longer SPs were comparatively few, especially within tone units. In the FP category, unit FPs occurred more often than brief but both types were rare.

Table 8:12 shows the overall distribution of SPs and FPs in relation to sentence, clause, phrase, and word transitions. The following general tendencies could be observed:

Ps between SENTENCES were unit or longer and occurred between TUs.
Ps between CLAUSES were generally brief and occurred between TUs.
Ps between CLAUSE ELEMENTS were brief and occurred between TUs.
Ps between WORDS in phrases were brief and occurred within TUs.
Ps with REFORMULATIONS were unit or brief and occurred between TUs.

Separation of sentences by means of pauses was mainly achieved with unit SPs; double and treble SPs were rare in any other position. Unit SPs were more often found between clause elements than between clauses, while brief SPs occurred chiefly between elements of clause structure.

FPs were rare in this position. Neither unit nor brief FPs cooccurred with sentence or clause boundaries. But it is possible to detect a slight tendency for unit FPs to occur between clause elements and in reformulations and for brief FPs to occur between lexical words (cf Maclay & Osgood 1959).

Crystal (1969:170) found that over 60% of all pauses occurred between clauses or elements of clause structure, but according to this study, the majority of the pauses occurred between clause elements and not between clauses. Maclay & Osgood (1959) observed that FPs tend to occur at phrase

**Table 8:12.** Pauses between and within (underlined) TUs in relation to syntactic components in text S.12.6.

| | SILENT | | | | | | | | FILLED | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | • | | – | | --/--- | | Total | | [ə:m] | | [əm] | | Total | |
| *Between sentences* | 82 | | 111 | | 61 | | 254 | | | | | | | |
| *Between clauses* | | | | | | | | | | | | | | |
| coordinate | 23 | | 2 | <u>1</u> | | | | | | | | | | |
| subordinate | | | | | | | | | | | | | | |
| relative | 9 | | 5 | <u>1</u> | | | | | | | | | | |
| other+conj | 14 | <u>6</u> | 4 | | | | | | | | | | | |
| -conj | 4 | | 1 | | 1 | | 63 | <u>8</u> | | | | | | |
| ellipsis/non-finite | 10 | | 3 | | | | 13 | | | | | | | |
| *Within clause* | | | | | | | | | | | | | | |
| A-A | 19 | <u>1</u> | 5 | <u>2</u> | 1 | | | | | | | | | |
| A-0 | 2 | | 1 | | | | | | | | | | | |
| A-S | 23 | <u>4</u> | 5 | | | | | | | | | | | |
| A-V | 9 | <u>3</u> | 2 | | | | | | | | | <u>1</u> | | |
| O-A | 17 | <u>3</u> | 1 | | 1 | | | | | | | | | |
| O-O | 6 | <u>2</u> | 1 | | | | | | | | | | | |
| O-S/V | 4 | <u>1</u> | 1 | <u>1</u> | | | | | | | | | | |
| S-A | 6 | <u>1</u> | | | | | | | | | | | | |
| S-O | | <u>1</u> | 1 | | | | | | | | | | | |
| S-V | 25 | <u>8</u> | 9 | | | | | | | | | | | |
| V-A | 14 | <u>3</u> | 4 | | | | | | | | | | | |
| V-C | 1 | <u>4</u> | 2 | | | <u>1</u> | | | | | | <u>1</u> | | |
| V-O | 9 | <u>9</u> | 1 | <u>2</u> | | | | | | | | | | |
| V-S | 1 | | 2 | | 1 | | | | | | | | | |
| V-V | 1 | <u>1</u> | 1 | | | | | | | | | | | |
| A-postmod | 6 | | 1 | | | | | | | | | | | |
| C | 4 | | 2 | | | | | | | | | | | |
| O | 8 | | 1 | | | | | | | | | | | |
| S | 7 | | 1 | | | | | | | | | | | |
| A-app | | | 3 | | | | | | | | | | | |
| C | 4 | | 2 | | | | | | | | | | | |
| O | 3 | <u>1</u> | | <u>1</u> | | | | | | | | | | |
| V-ag | 3 | | | | | | 221 | <u>49</u> | | | | | | |
| conj-clause | | <u>5</u> | | <u>1</u> | | | | | | | | | | |
| conj P | 14 | | 4 | | | | | | | <u>1</u> | 3 | | | |
| other | 17 | | 3 | | | | 38 | <u>6</u> | 1 | | | | 4 | <u>3</u> |
| *Within phrase* | | | | | | | | | | | | | | |
| APH | 5 | <u>5</u> | | <u>1</u> | | | | | | | | | | |
| JPH | | <u>1</u> | | | | | | | | | | | | |
| NPH | 11 | <u>24</u> | 1 | <u>2</u> | | | | | | <u>3</u> | | | | |
| PPH | 2 | <u>28</u> | 1 | | | | | | | <u>1</u> | | | | |
| VPH | 5 | <u>13</u> | 1 | | | | 26 | <u>74</u> | | | | | | <u>4</u> |
| *Reform* | 4 | <u>4</u> | 2 | | 1 | <u>2</u> | | | | <u>1</u> | | | | <u>1</u> |
| other | 15 | <u>1</u> | 7 | | 1 | | 30 | <u>7</u> | | | | | | |
| TOTAL | 387 | <u>129</u> | 191 | <u>12</u> | 67 | <u>3</u> | 645 | <u>144</u> | 1 | <u>6</u> | 3 | <u>2</u> | 4 | <u>8</u> |

boundaries and not within phrases. These observations were not contradicted by my findings.

### 8.10.1 Pauses between sentences

Pauses between sentences were all of the SP type, and sentence boundaries with a pause were coterminous with tone unit boundaries:

(20) and ||he made ΔSefton PÁRK■ • ||his 'English ΔHÒME■ – "||great ALTERÀTIONS were MÁDE■ – (S.12.6:587-589)

The majority of the pauses at sentence boundaries were unit or longer:

| | |
|---|---|
| Brief | 32% |
| Unit | 44% |
| Double | 14% |
| Treble | 10% |

Note that a number of strings defined as sentences are single-clause sentences. It should also be observed that a large number of the double and treble SPs served not only as sentence demarcators but also as discourse markers. FPs served neither of these functions.

### 8.10.2 Pauses between clauses

Clause and tone unit boundaries were nearly always coterminous. Consequently, pauses separating clauses were generally found between tone units. 79% were brief SPs and 21% were unit SPs, ie usually shorter between clauses than between sentences. FPs did not occur between clauses (see Table 8:12).

In the eight cases where a clause juncture with a pause occurred within the tone unit, the subordinate clause served as a constituent of the superordinate clause, eg direct object:

(21) that they ||think • that 'I am 'one as ΔWÈLL■ – (S.12.6:264)

or it was embedded as a hedge in the superordinate clause:

(22) ||where 'we were 'had • what you ||might call • Δrunning WÀTER■ (S.12.6:165)

One characteristic feature of the spoken language is its frequent use of coordinating conjunctions as links between sentences in a narrative. Compare (23) and (24):

(23) he ||certainly 'didn't come BÀCK■ – ||but the ∆PRÒPER▷TY■ • was ||taken ÒVER■ ||by the CUSTǑDIAN■ of ||Enemy PRÒPERTY■ (...) and "||all the furniture • was SÒLD■ (S.12.6:637-642)

(24) ||[ði:] ∆middle TÈNT■ ||was a great ∆DRǏNKING 'tent■ ||AND■ – in ||THÒSE DÁYS■ I ||think ÈVERYBODY■ – ||did more ∆heavy DRÍNKING■ (S.12.6:532-537)

*And* in (23) is used as an ordinary coordinating conjunction. The *and*-clause cannot stand on its own as a sentence, nor can *and* be left out without a noticeable effect. In this case *and* is preceded by the pause, which is the typical order for pauses and coordinating conjunctions (cf Table 8:13). By contrast, *and* in (24), which constitutes a separate tone unit and is followed by a pause, serves as a link in the discourse very much like the initiator *now* or the particle *well*. One indication that this *and* does not function as an ordinary coordinating conjunction is that it can easily be left out in the same way as *now* and *well* in a similar position.

The pause preceded the conjunction also in the cases of coordination and subordination where clauses met within the tone unit. When sentence and clause boundaries were coterminous with tone unit boundaries (as was generally the case), the pause occurred as in Table 8:13, which shows that the tendency for pauses to precede the conjunction was stronger at clause junctures (94%) than at sentence junctures (82%).

**Table 8:13.** Pauses and conjunctions cooccurring with sentence and clause boundaries following a tone unit boundary.

| | P CONJ | | P CONJ | | CONJ P | | TOTAL |
|---|---|---|---|---|---|---|---|
| SENTENCE / SENTENCE | 60 | 82% | 11 | 15% | 2 | 3% | 73 |
| CLAUSE / CLAUSE | | | | | | | |
| coord | 28 | | 1 | | | | 29 |
| subord | 13 | | | | 3 | | 16 |
| SUBTOTAL | 41 | 94% | 1 | 1% | 3 | 5% | 45 |
| TOTAL | 101 | 86% | 12 | 10% | 5 | 2% | 118 |

**Table 8:14.** Pauses and conjunctions between tone units at sentence boundaries.

| | P CONJ | P CONJ P | CONJ P | TOTAL |
|---|---|---|---|---|
| COORDINATORS | | | | |
| *and* | 47 | 10 | 1 | 58 |
| *but* | 7 | 1 | | 8 |
| *so* | 5 | | | 5 |
| *because* | 1 | | 1 | 2 |
| TOTAL | 60 | 11 | 2 | 73 |

Here are some examples with both coordinating and subordinating conjunctions. First, in (25) and (26) pause + conjunction between tone units:

(25) – ||funnily e'nough my ▵FATHER■ · ||WENT to the 'same SCHOOL■ · and ||he was 'one of the 'first ▵PUPILS■ – (S.12.6:13-15)

(26) – it was ||only AFTER■ ||WORLD War TWO■ · that ||WE■ ||REALLY■ the ||PEOPLE 'in our VILLAGE■ · ||HAD■ – a ||little to SAY■ · (S.12.6:372-378)

Then, conjunction + pause between tone units (27):

(27) BE||CAUSE■ · on ||bath · NIGHTS■ (S.12.6:184-185)

And within tone units (28):

(28) ||and · "▵LOOKING■ ||at the ▷< lie of the ▵LAND■ (S.12.6:1051-1052)

Table 8:14 shows that *and* with a pause served to initiate a sentence more often than any other conjunction.

### 8.10.3 Pauses between clause elements

The largest proportion of pauses separating clause elements was found between tone units. As a matter of fact, 33% of the total number of pauses were found between tone units (and between clause elements) and 7% were found within tone units (and between clause elements). SPs other than brief and unit were rare in both positions (Table 8:12). Here are two examples, in (29) with a pause between verb and object:

(29) of ‖course it would ▵ÌNTEREST – MÉ {RÉALLY■ }■ (S.12.6:617)

in (30) between subject and verb:

(30) ‖and per▵haps Δsome of 'you • could RE▵MÈMBER■ (S.12.6:230)

The following examples illustrate pauses between tone units. In (31) the pause occurs between adverbial and subject and in (32) between two adverbials:

(31) ‖in 'nineteen ▵TWÈLVE■ – he ‖WÉNT■ – (S.12.6:518-519)

(32) he ‖STÀYED there one NÍGHT■ – on his ‖way to be ▵ÈXECUTED■ • (S.12.6:1129-1130)

### 8.10.4 Pauses between words in phrases

Words in phrases were also separated by SPs, generally within the tone unit. Most of them occurred within noun phrases as in (33) and prepositional phrases (see Table 8:12).

(33) ‖his – Δbeautiful ▵copperplate WRÍTING■ (S.12.6:26)

### 8.10.5 Pauses and reformulations

Only 14 instances of reformulation involved a pause (or pauses). There was no tendency for some types of reformulation to occur within the tone unit and for others just after a tone unit boundary (Table 8:12). Here are examples of some different types:

(34) ‖was – – ‖had ‖CONΔNĔCTION■ (S.12.6:700)

(35) she ‖made some Δvery • ‖knitted some Δvery itchy VÈSTS■ (S.12.6:90)

(36) during [ði:] • ▵Thomas Gray ▵[fest • ə:] CENΔTÈNARY■ (S.12.6:819)

(37) the ‖large HÓUSES■ ‖in this [vi] ‖in this VÍLLAGE■ • ‖ÀLL the 'large MÁNSION«S»■ (S.12.6:335-337)

(38) ‖when – HÈ■ – – when Mr ▵Fortune RE▵TÍRED■ (S.12.6:981-982)

(39) ‖and we LÍVED■ • we were a ‖FÀMILY of ▵FÍVE■ • (S.12.6:38-39)

(40) ||WHI→CH■ - [ði:] "||HI∨GHLAND di'vision■ - was ||stationed HE↘RE■ (S.12.6:645-647)

(41) «it» and ||thing • the ||Women's △I↘NSTITUTE {and ||things like THA↘T■ }■ (S.12.6:998)

(42) ||only [ði:] • ||possibly 'Lord«s» of the △MA↘NOR■ - (S.12.6:1147)

(43) ||and 'when we 'were -- ||I was 'in the △CHO↘IR■ - (S.12.6:284)

### 8.10.6 Performance units

Performance units can vary a great deal in length, as in (44):

(44) - the ||GRO→UNDS■ were ||REO↗RGANIZED■ • ||and he △INTRODU↘CED■ ||into the △VI↘LLAGE■ • the most ||modern 'farm △BU∨ILDINGS■ that - ||in that PE↗RIOD■ were ||rather U△NI↘QUE■ (S.12.6:590-596)

With brief SPs regarded as minimal delimiters, the performance units in (44) vary from two to six words. If only unit SPs are considered, they vary from two to 13 words. This should be compared with the average number of words per pause, which is 6.3 in this text and 9.4 in the larger subcorpus consisting of ten texts (see p 214). In (45) the performance units vary in length from one to 13 words:

(45) - and ||looking 'through my BINO↘CULARS one ▻ day■ I ||saw on the △{O↘PPOSITE} △MO↘UNTAIN■ - a ||MA∨N■ || WO∨RKING■ ||on ['ei] ||one of THE↗SE■ • ||dry 'stone WA∨LLS■ • di||viding • the BO↘UNDARIES■ • (S.12.6:658-664)

In both (44) and (45) unit SPs separate old from new information. Once the speaker is on the right track, brief SPs serve as demarcators. The somewhat stilted word order in these extracts seems to invite a pause:

> and he introduced into the village . the most modern farm buildings that - in that period were rather unique
>
> I saw on the opposite mountain - a man

The use of nonfinite clauses is more characteristic of writing than of speech: *looking through my binoculars, dividing the boundaries.*

In quite a few cases there was a pause between clause constituents, eg between subject and verb (*the grounds - were reorganized*), and between verb and object (*dividing* · *the boundaries*). These pauses are difficult to explain in terms of linguistic demarcators. One often gets the impression, when listening to the recording, that the speaker inserts a pause before or after a particular word or string of words to obtain a certain emphatic effect rather than using pauses as linguistic demarcators (and for breathing): 'the speaker prosodically empathizes with the hearer' (Quirk et al 1985:1444).

In order to identify a performance unit it was necessary to verify to what extent sentences and clauses were demarcated by a pause, and by what type of pause, and also to examine whether they constituted pause-free units (see Tables 8:15 and 16). Table 8:15 indicates that 253 out of 293 sentences (86%) had an initiating pause; 82 out of 171 finite clauses (48%) had an initiating pause. Perhaps the most striking finding is that unit or longer SPs were more than twice as frequent as brief SPs, not only between sentences and clauses but also between clause constituents (Table 8:16). But note that, in the data as a whole, brief SPs were more frequent than unit or longer SPs between clauses and clause elements.

Table 8:17 gives the percentage of pauses (all types) at sentence and clause boundaries, between clause and phrase elements, and in reformulations. The distribution in the syntactic structure was different between and within the tone units. Among the total occurrence of pauses between tone units those separating sentences and those separating elements of clause structure made up an almost equal percentage (40% and 41%), while the percentage of pauses separating clauses was small (12%), and pauses between phrase elements (single words) constituted the smallest group (4%). Within the tone unit the largest percentage consisted of pauses between words in phrases (51%); the

**Table 8:15**. Sentences and clauses with and without an initiating pause.

| GRAMMATICAL UNIT | + PAUSE | | - PAUSE | | TOTAL |
|---|---|---|---|---|---|
| sentence/sentence | 253 | 86% | 40 | 14% | 293 |
| clause/clause | 82 | 48% | 89 | 52% | 171 |
| TOTAL | 335 | 72% | 129 | 28% | 464 |

**Table 8:16**. 'Sentences' initiated by a pause.

| TYPE of initiating PAUSE | FLUENT SENTENCES S = one clause | S = more than one clause | NON-FLUENT SENTENCES P between clauses | P between constituents | TOTAL |
|---|---|---|---|---|---|
| UNIT or longer | 24 | 25 | 30 | 92 | 171 |
| BRIEF | 19 | 16 | 9 | 38 | 82 |
| TOTAL | 43 | 41 | 39 | 130 | 253 |

proportion of pauses between clause elements was fairly high (38%), while pauses at clause boundaries were rare (6%), and pauses at sentence boundaries were non-existent.

The reason why pauses within tone units seldom preceded a clause and never a sentence is of course that sentence boundaries were always, and clause boundaries nearly always, coterminous with tone unit boundaries.

**Table 8:17**. Pauses within and between tone units in the syntactic hierarchy, including reformulations.

| | BETWEEN TUs | WITHIN TUs |
|---|---|---|
| Between sentences | 40% | 0% |
| Between clauses | 12% | 6% |
| Between phrases | 41% | 38% |
| Between words | 4% | 51% |
| Between reformulations | 3% | 5% |

### 8.10.7 Summing up

- 32% of all SPs occurred between sentences; half of them were unit and longer (Table 8:12).
- The proportion of pauses between clauses was only 11% (Table 8:12).
- 39% of all SPs separated phrases as compared with 2% of all FPs; especially sequences consisting of phrases serving as S and V, V and O, A and A (Table 8:12).
- 12% SPs and 1% FPs separated words in phrases; such pauses generally occurred within the tone unit (Table 8:12).
- 81% of all SPs and FPs occurred between tone units; brief FPs were always found within the tone unit; brief SPs in 35% of the cases (Table 8:12).
- More sentences than clauses were initiated by a pause + a conjunction, and coordination exceeded subordination; conjunction + P order was rare but slightly more common with subordinated clauses than anywhere else (Table 8:13).
- *And* dominated as a coordinator initiating sentences (Table 8:14).
- Performance units (separated by a pause on both sides) varied in length from one to 13 words, ie they were either shorter or longer than a clause.

No safe rules for pause assignment can be based on the results of this study alone. First, one single text has been examined; second, the definitions of sentence and clause as used in spoken discourse are fairly vague; third, although the study shows the ratio of pauses to phrases and words, it does not show the corresponding ratio of absent pauses.

## 8.11 Pauses and verbal fillers as discourse markers

### 8.11.1 Stage markers

Pauses and verbal fillers can be used unconsciously, as when they occur with breathing or as hesitation signals. They can also be used consciously, for example as structural markers or as a means for the speaker to manipulate the listener and save face. In this section I shall report on pauses and verbal fillers used as organizational and interactional devices.

In addition to functioning as linguistic markers in the syntactic hierarchy (Section 8.10), SPs typically serve to 'mark stages' in the discourse (Labov & Fanshel 1977:156) and to organize the talk into 'paragraphs' in the narrative structure (Chafe 1987:44). This is illustrated in Figure 8:3, which is a

**Figure 8:3**. Pauses as discourse markers. (w = text)

preface NOW■ - let's go ||over to some interesting △BUILDINGS■
-

TOPIC 1 △STOKE House
wwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwww
wwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwww
wwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwww
wwwwwwwww it's last OWNER■
---

TOPIC 2 the ||WATER SUPPLY■ • ||to our HOUSE■ ||was from a △WELL■
wwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwww
wwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwww
digging the △WELL■
--

subtopic ||my FATHER■ ||was a water DI△VINER■ •
wwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwww
wwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwww
wwwwww △running WATER■
---(laughter)

subtopic BE||CAUSE■ •on ||bath • NIGHTS■ we ||had to light the △COPPER■
wwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwww
wwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwww
wwwwwww ||in the same WATER■
-- (laughter)

TOPIC 3 SUNDAYS■ wwwwwwwwwwwwwwwwwwwwwww "△CHURCHES■

contracted version of a longer monologue-sequence intended to show how the speaker divides the narrative into a number of paragraphs and subparagraphs, each dealing with a particular topic or subtopic that is part of the main theme. It was possible to identify a large number of such topic and subtopic

**Table 8:18**. SPs separating topics and subtopics.

| TYPE OF PARAGRAPH | TYPE OF PAUSE Treble | Double | Unit | TOTAL |
|---|---|---|---|---|
| topic | 15 | 7 | 9 | 31 |
| subtopic | 10 (5) | 6 (2) | 5 (3) | 21 (10) |
| TOTAL | 25 (5) | 13 (2) | 14 (3) | 52 (10) |

paragraphs. Pauses between topic paragraphs were found to be longer than those separating subtopic paragraphs. The distribution is shown in Table 8:18 (where the figures in brackets indicate that subtopic pauses coincided with topic-boundary pauses). The following characteristics emerged:

- Treble SPs dominated followed by unit SPs for both categories of topic; double SPs were the least common type.
- 71% of the topic paragraphs and 76% of the subtopic paragraphs ended with an SP longer than unit.
- More treble than double SPs and more double than unit SPs separated topics; the same tendency was observed for subtopics.

Topic transitions were also reflected in changes in the pitch contour, indicated by boosters (see Chapter 7) in our analysis, an area that I shall not go into, however.

### 8.11.2 Prefaces

Almost half of the topic paragraphs were initiated by a preface (which introduces the topic), all but three accompanied by an SP, usually longer than unit. A preface can be long (46) or short (47):

(46) I'm ‖now GOING■ to ‖talk ABOUT■ • the ‖different HOUSES■ – ‖and ‖IN the 'village■ ‖and I'm △going to 'start OFF■ with ‖Wexham SPRINGS■ – ‖which is the △HOME TODAY■ ‖of Ce'ment and △CONCRETE As'soci'ation■ – ‖which has GOT■ a ‖special INTEREST • to ME■ ––– (S.12.6:390-399)

(47) we ‖now go OVER■ • to ‖Stoke COURT■ –– (S.12.6:697-98)

In some cases an initiator alone served to introduce the paragraph, as in (48):

(48) ‖NOW■ – ‖AFTER • the ALLUYSONS■ (S.12.6:734)

Conjunctions (especially *and*, *but* and *because)* followed by a pause and the conjunct *so* were often used as links between paragraphs:

(49) ‖BUT [ə]■ we ‖now 'go to ▵Stoke "△HOUSE■ (S.12.6:1165-1166)

A paragraph may also be initiated by the 'fronting' of the new topic:

(50) ‖LARCHMOOR■ --- of ‖course be△fore the present 'school was BUILT■ there ‖stood a a ‖quite a big (...) (S.12.6:952-954)

The following tendencies could be observed:

- Topics were introduced by a preface plus a unit or longer SP.
- The preface was either introduced by an initiator (eg *now*) followed by an SP or consisted of nothing but the initiator followed by an SP.

### 8.11.3 Framing

The previous section was devoted to discourse markers in the monologue. In this section I will comment briefly on discourse markers in the dialogues. Many dialogues include narrative sections which resemble monologues and where SPs and VFs, in combination with SPs, serve as 'stage markers' in much the same way. Compare (51) and (52), where *anyway* + SP brings the discourse 'back to order'. Both extracts illustrate that the thread of discourse was momentarily lost:

(51) *but* ‖that didn't ▵HAPPEN■ until ‖LONG 'after [ði:]■ -• [ə:] -- ‖British and 'French and A▵merican – ARMIES■ had ‖really sort of --- ‖anyway I'm ▵SORRY■ I was DI‖GRESSING■ • but ‖what I 'mean IS■ -- the ‖German 'General STAFF■ was ‖no – was was ‖very IM▵PRESSIVE■ in its ‖HEYDAY■ (S.2.3:362-370)

(52) A: ‖that ▻brought △HER a'bout■ • ‖her and △what you MIGHT ▻say■ "‖HER little outlook on ▻life■

b: oh yes *«I can» understand her*

A: *||ÂNY'WAY■ •* "||ÂNY'WAY■ - the ||next MÓRNING■ -- ||[sʌ] ||SǑMEHOW or 'other■ I ||hadn't 'got ▵any - BÙSINESS to 'do■ (S.1.14:484-492)

But note that in (52) *anyway* (in a separate tone unit) + SP has an additional, interactive function: it is also used as a device for preventing the other party from taking over the turn (the asterisks indicate simultaneous speech). *Anyway* in (51), which is preceded by a treble SP, does not have that function. *However* + SP can be used in a similar way, indicating 'let's go on'.

The function of *now* + SP is slightly different: it only points forward and would probably never be used as a turntaker/turnholder:

(53) A: ||but [?]- ||then this 'other this ||pædia∆trician 'showed me a ▵PÌCTURE■ of a ||young 'kid who had 'bow ∆LÈGS■ and said ||what's the ▹DIAGA∆NÒSIS■ ||so I said "∆RÌCKETS■ (...)

a: renal rickets

A: yeah it's ||called 'renal ▵RÌCKETS■ (...)

A: ||NŌW■ -- ||then he said "∆{ÁH} YÉS■ - «well ||then» I 'told him AN▵ÒTHER 'cause of 'rickets■ (S.2.9:127-155)

### 8.11.4 Summing up

- SPs alone may serve as stage markers to demarcate the preface from the narrative proper and to indicate 'end of topic' and topic transition.
- FPs, alone or in combinations, signal hesitation.
- VFs, accompanied by an SP (or alone), indicate either 'resumption of lost topic and continuation' or the beginning of a new topic.

## 8.12 Discourse interaction

In addition to the functions accounted for so far, pauses and fillers are used as interactional and social devices. What pauses and fillers do in the interaction is not only a function of the way they are realized but also of their position (cf Crystal & Davy 1975:92ff and Chapter 5:

(54) A: ||think I'm 'on the 'wrong ▵MÀP■ ---

B: [ə:m m] but ||that's as a'bout as ▵near as 'you can ▵hit it ǑFF■ -- ||and it's [ə] --- it's ||NÒT a 'road {you could ||really ADVÌSE■ }■ --- «to» ||somebody who didn't KNÓW it■ --- [m] «3 sylls» ‖ǑTHERWISE■ you've ‖got to do

ΔÀLDERSHOT■ ---

A: ‖YÈAH■ --- well in ‖THǍT ⊳case■ (S.1.11:781-788)

Example (54), a short extract from a rather special situation with two persons reading a map, contains eight long SPs which occur alone or in combinations. The pauses can have different functions:

turntaking: preceded by an FP

*[ə:m] --- but*

turnholding: signalling hesitation as when coinciding with repetition

*and it's [ə] --- it's*

turnyielding:

*Aldershot ---*

The example illustrates that pauses and fillers can be produced both within a speaker's turn and at speaker shifts. The use of pauses and fillers is both situation-specific and speaker-specific. Especially VFs are for instance less likely to occur in a formal than an informal situation, and some speakers use pauses and/or fillers much more frequently than others.

### 8.12.1 Turntaking

If a person remains silent too long at a potential speaker shift he will never get the turn. If he wants to take the turn he has to say something, even if he has not yet made up his mind exactly what to say. One way of starting is by a filled pause or a verbal filler, for example: *-[ə:m]* or *well* - -, and more items can be added to fill for time:

[əm]
[ə:m] well
[ə:m] well you know
[ə:m] well you know I mean

Note that additional items imply additional functions; what started as pure hesitation (-) develops into stalling (*well*) and interpersonal activity (*you know I mean*) before the speaker arrives at his real message. The order of the items and possible combinations are not fixed, but some patterns are much more likely than others. *Well*, for instance, should come early.

The beginning of a turn is generally where the global planning of an utterance takes place, while the local word-by-word planning is done within the turn (cf Clark & Clark 1977:248ff for 'skeleton' and 'constituent' planning). So even if hesitation phenomena can occur almost anywhere, complex hesitation 'strings' are more likely to occur turn-initially, as in (55), but there are numerous counter-examples:

(55) B: I ‖think they've got quite a good O△PINION of him■ -

A: ‖well [ə] ▵I △I have TOO■

B: ‖[m]■

A: [ə:m] -- «‖well I mean» • the ‖way these chaps △GO■ (S.1.1:592-595)

### 8.12.2 Turnholding

A speaker who wants to keep his turn cannot afford to remain silent for long, unless he is in a position to prevent the other party from breaking in anyway. This is where FPs and VFs, or combinations of both, help him to gain time:

(56) A: (...) [ə] but I ‖heard it •[ə] mentioned by somebody ELSE■ - I ‖think △WATT■ - I'm ‖not SURE■ *-* [ə:m] -- ‖and [ə:m]

B: *‖[m][hm]■ *

>A: you KNOW■ - ‖if △this is ▵ALSO COME■ ‖from [ə] - NIGHTINGALE (...) (S.1.1:265-271)

The hedge *sort of*, which was never found at the end of a turn in this data, frequently occurred in turn-medial position with various other functions besides acting as a turnholder:

(57) ‖well I △don't 'think • 'it's ‖«sort of a» • a com▵plete CON▵CLUSION■ you're sort of ‖left with the -- you ‖sort of [ə:m] - it's ‖sort [ə?] an ▵end to a ▵story in a ▵WAY■ • you can ‖just im'agine ▻these ▻things ▻going ON■ it ‖sort of △winds UP■ it's [ə:m] • ‖rather an △ARTI△FICIAL • ‖{[du:nei'mã]■ }■ ‖rather 'like [?] 'one of [ə:m] 'MOLI△ERE'S 'plays■ ‖where they ▻sort of • bring IN a • a man at the END to■ • to ‖finish 'everything OFF■ ‖round it △OFF■ (S.3.5:143-151)

In (57), *sort of* with an SP or an FP, or both, occurs in no less than six places in the same speaker's turn. It is obvious that he has some difficulty in putting his thoughts into words. Other indications of hesitant behaviour are

repetitions: *sort of a • a complete conclusion*
*the end to • to*

and reformulations: *you're sort of left with the -- sort of [ə:m] - it's sort [ə] an end to a story in a way*
*it's [ə:m] • rather an artificial*

However, *sort of* does not only indicate hesitation. It may also have a softening effect:

*a sort of simple, incredibly mixed up sort of chap*

it may be used as a hedge, in which case it generally modifies another vague word or expression:

*sort of stuff / thing / general course*

it may serve to indicate approximation:

*sort of five, six million*

or it may have a pejorative effect (see further Aijmer 1984):

*a sort of dinner*

In the present data, *I mean* and also *you know* and *you see* were more often found within the turn than in turn-initial and turn-final position. And *well* occurred unexpectedly often within the turn. Obviously, a great deal of planning takes place within the turn. Turn length is of course an important variable; if the turn is long, the speaker is likely to undertake some new global planning within the turn.

Besides signalling hesitation, the VF categories are inherently different. *Anyway*, *however*, and *now* are message-oriented and connected with the organization of the discourse, while *I mean*, *you know*, and *you see* are mainly interpersonal and appeal for understanding and feedback.

*Sort of* is also message-oriented but, unlike the other verbal fillers, often directly involved in the syntactic process instead of standing 'outside'. It can be described in syntactic terms as a modifier of a preceding or following content word or phrase.

The staller *well* is the most 'neutral' verbal filler, very often equivalent to the filled pause *ə:m*. The fact that the softeners *I mean*, *you know* and *you see* were twice as frequent as the staller *well* seems to indicate that the social aspect with the speaker appealing to the listener for understanding and feedback is more pronounced in conversation than just the need to gain time for planning.

### 8.12.3 Turnyielding

SPs are of course the most typical turnyielders. Whether they should be attributed to the speaker who yields the turn or to the one who takes over is difficult to determine from reading a transcript. However, one might speculate that SPs serve as turnyielders more often in spontaneous conversation than, for instance, unprepared interviews where the addressee probably needs more time. The FP *[ə:m]* followed by an SP may occur at the end of a turn, however, as a speaker's last effort to keep the turn when he does not know how to continue, as in (58):

(58) [ə:m] --- it ||seems to ME■ that - [ə:m] --- (S.1.1:627-628)

VFs like *you know* and *you see* (often with a rising tone) cooccurring with SPs add a social dimension by not only appealing for understanding but also inviting feedback, preferably agreement, or just a minimal response like *mhm*, or laughter:

(59) A: ||which is «a» GREAT help■ - and ||then he says △course «if» you △don't
UNDERSTAND this■ - this ||subject's △not for YOU■ • ( • laughs) you ||KNOW■
B: ( - laughs) - (S.1.6:919-923)

The 'inviting force' is affected not only by lexical choice and choice of tone but also by the position of the SP before or after VF. Compare eg (a-d) below (note that VF in (b) and (d) would occur in a separate tone unit):

(a) this is exactly what he did you know --
(b) this is exactly what he did -- you know
(c) this is exactly what he did isn't it --
(d) this is exactly what he did -- isn't it

It seems that the urge for the listener to provide feedback is more pronounced in (b) than in (a), and in (d) than in (c), and that the degree of 'questionness' increases from (a) to (d).

### 8.12.4 Summing up

- Both pauses and VFs were used as hesitators.
- More treble than double SPs separated both topic and subtopic paragraphs. Unit SPs were somewhat more common than double SPs between topic paragraphs whereas double SPs were somewhat more common than unit SPs between subtopic paragraphs.
- Double or longer SPs separated topic and subtopic paragraphs in the monologue.
- Approximately half of the topic paragraphs were introduced by a preface, usually followed by a unit SP.
- Both pauses and VFs served as interactional devices in the dialogue: as turntakers, turnholders, and turnyielders.
- VFs with or without a pause served as social devices and as markers of discourse organization.

9

# Adverbial commas and prosodic segmentation

*Anna-Brita Stenström*

## 9.1 Introduction

One important aim of the TESS project has been to suggest rules for prosodic segmentation of written English on the basis of the prosodic patterns typical of genuine English speech as manifested in a corpus of spoken English. Obviously, the segmentation procedure to be used in speech synthesis would be simplified and speeded up if it was possible to find, in written texts, reliable cues to prosodic correspondences in speech. One such cue is punctuation, judging by Quirk et al where it is stated that there is a direct relation between speech and writing and 'also (broadly) between the prosodic features of speech and the punctuation devices of writing' (1985:1443). In this chapter I will investigate the evidence of this relation in the London-Lund Corpus and its usefulness in making predictions about segmentation.

It is assumed that all punctuation marks indicate a prosodic break. Therefore, we have used rules to convert these marks (. ? ! : ; , -) automatically into a tone unit boundary, unless there are specific constraints (pp 303ff). The comma is not a reliable customer, however:

> The comma indicates the smallest interruption in continuity of thought or sentence structure. There are a few rules governing its use that have become almost obligatory. Aside from these, the use of the comma is mainly a matter of good judgment, with ease of reading as the end in view. (*A Manual of Style* 1969:108)

A comma between clauses can usually be taken as a fairly safe boundary marker (cf Altenberg 1987a:52). Also, sentence adverbials are separated 'from the rest of the clause by intonation boundaries in speech or by commas in writing' (Quirk et al 1985:52). Is this comma also a safe boundary marker, and what about the other adverbials? The specific question to be discussed in this chapter is the following: 'If a particular adverbial is typically preceded and/or followed by a comma in writing, will it also typically be preceded and/or followed by a tone unit boundary in speech (under similar circumstances)?'

Meyer, who studied punctuation practice in the Brown Corpus, found that the various prosodic patterns that are typical of adverbials in speech were mirrored by punctuation in the written material in the following way (1986:73):

- Adverbials that always constituted separate tone units were usually punctuated.
- Adverbials that optionally constituted separate tone units were either punctuated or unpunctuated.
- Adverbials that generally did not constitute separate tone units were usually not punctuated.

If such agreement exists, it should be possible to take punctuation as a fairly reliable cue for prosodic breaks.

## 9.2 The study

Starting from the assumption that there is such agreement, I examined the correlation between commas setting off certain adverbials in writing (LOB) and the prosodic separation of the corresponding adverbials in speech (LLC). The main reason for choosing adverbials was that some such categories are typically marked off by a comma, notably disjuncts, eg *actually* and *obviously*, and conjuncts, eg *however* and *therefore* (Quirk et al 1985:620, 927). Also, closed-category adverbials were easily retrievable in the concordance versions of the corpora.

The complexity of adverbial prosody has been demonstrated by Allerton & Cruttenden (1974, 1976, 1978), who studied the correlation between intonation groupings and syntactic-semantic classes. They did not examine a genuine spoken corpus, but used randomly collected data which they tested in various ways and checked against the intuitions of colleagues. I often find their results, which have been scrutinized by native speakers and undergone thorough testing, more convincing than mine, which are affected by situational parameters out of my control. I have therefore used their proposals both as a support and as a guiding norm in doubtful cases when suggesting what seems to be the most adequate rules for the automatic assignment of tone unit boundaries .

In order to find out to what extent adverbial commas in writing are matched by tone unit boundaries in speech I selected a number of frequent adverbials that are usually punctuated and compared the occurrence of commas in LOB with the occurrence of tone unit boundaries in LLC. The reason for choosing LOB rather than the Brown Corpus was simply that, like LLC, LOB represents British English while the Brown Corpus consists of American English. Whether the punctuation rules are the same or different in British and American English was not taken into account.

The study includes the following adverbial categories and involves 43 different items, such as the following (for a complete list, see Figure 9:5):

| | | |
|---|---|---|
| AA | adjuncts | *again, still* |
| AC | conjuncts | *anyway, for instance* |
| AD | disjuncts | *frankly, generally* |
| AQ | discourse items | *is it were, please* |
| AS | subjuncts | *indeed, obviously* |

Adverbials can occur in initial (I), medial (M), and final (E) sentence position. The positions are defined as follows (in a simplified version of the definitions in Quirk et al 1985:490ff):

I before all obligatory clause elements (including cases where another adverbial precedes, eg *well now*); initial adverbials may be followed by a comma;

M between the first and the last obligatory clause element; medial adverbials may be preceded and/or followed by a comma;

E after all obligatory clause elements; final adverbials may be preceded by a comma.

The position of an adverbial does not only affect its punctuation/prosody but also its grammatical function. For example *now*, which is usually a transitional conjunct if placed at the beginning of a sentence, is a time adjunct in medial or final position; and *clearly*, which normally serves as a content disjunct in initial position, generally converts into a process adjunct in final position:

(1) *Now* would you believe it? He's in China *now*.

(2) *Clearly* I don't believe it. I've made the point quite *clearly*.

**Table 9:1.** Frequencies of the selected items in LOB and LLC (in alphabetical order).

| ITEM | TAG | LOB | LLC |
|---|---|---|---|
| after all | ACcon | 17 | 37 |
| again | AAtim | 663 | 403 |
| again | ACcon | 14 | 36 |
| also | AClis | 944 | 238 |
| anyway | ACcon | 58 | 280 |
| apparently | ADcnt | 69 | 57 |
| as a matter of fact | ACrep | 8 | 15 |
| as it were | AQsof | 7 | 34 |
| basically | ADcnt | 12 | 33 |
| clearly | ADcnt | 23 | 17 |
| clearly | AApro | 101 | 17 |
| especially | ACapp | 167 | 10 |
| eventually | ACtra | 71 | 24 |
| finally | AClis | 166 | 28 |
| for example | ACapp | 126 | 49 |
| for instance | ACapp | 86 | 53 |
| fortunately | ADcnt | 34 | 16 |
| frankly | ADsty | 7 | 20 |
| generally | ADsty | 130 | 27 |
| honestly | ADsty | 1 | 15 |
| honestly | AApro | 6 | 2 |
| honestly | AQres | | 2 |
| however | AApro | 15 | |
| however | ACcon | 511 | 36 |
| however | ASint | 40 | 8 |
| in fact | ACrep | 146 | 465 |
| in particular | ACadd | 58 | 14 |
| indeed | ADcnt | 162 | 104 |
| indeed | ASint | 85 | 60 |
| indeed | AQres | 2 | 7 |
| instead | ACcon | 165 | 9 |
| maybe | ADcnt | 82 | 45 |
| moreover | ACadd | 55 | 3 |
| nevertheless | ACcon | 92 | 22 |

| | | | |
|---|---|---|---|
| now | AAtim | 1458 | 959 |
| now | ACcon | 31 | 452 |
| obviously | ASfoc | 124 | 169 |
| of course | ADcnt | 146 | 680 |
| on the other hand | ACenu | 78 | 36 |
| perhaps | ADcnt | 406 | 291 |
| please | AQpol | 94 | 117 |
| presumably | ADcnt | 46 | 50 |
| probably | ADcnt | 270 | 316 |
| so | ACres | 2413 | 1220 |
| still | ACcon | 10 | 13 |
| still | AAtim | 833 | 14 |
| surely | ADcnt | 126 | 43 |
| therefore | ACsum | 296 | 64 |
| thus | ACapp | 257 | 3 |
| too | AClis | 923 | 49 |
| unfortunately | ADcnt | 44 | 25 |
| yet | ACcon | 441 | 43 |

The frequency of the selected items in LOB and LLC is given in Table 9:1. The following general tendencies emerged. First, the distribution is not the same in spoken and written production: *however*, *too*, and *yet*, for instance, were much more frequent in LOB than in LLC, while there were fewer occurrences of *as a matter of fact*, *as it were*, *basically*, and *in fact* in LOB than in LLC. Second, the predominant function of an item was sometimes different in the two media: *now* occurred much more frequently in LOB, generally as a time adjunct, than in LLC where it was more often used as a transitional conjunct. Third, function is related to position: initial *again* and *now* were generally used as transitional conjuncts but served as time adjuncts when occurring later in the sentence.

Concentrating on the occurrence of commas and tone unit boundaries per adverbial item (orthographic word or multi-word), regardless of whether it served one or more than one function, I arrived at the results in Figure 9:1. It displays the percentage of comma per adverbial and position in LOB and the corresponding typical prosodic pattern per adverbial and position in LLC. The correlation between comma in LOB and tone unit boundary in LLC was fairly high for adverbials occurring at the beginning of a sentence but decreased with

**Figure 9:1.** Correlation between comma and TU boundary per adverbial; arrow (—>) indicates agreement 'corresponds to'; broken arrow (-/->) indicates disagreement 'does not correspond to').

| | | | | | LOB | | LLC |
|---|---|---|---|---|---|---|---|
| Initial position (I) | | | | | | | |
| AGREE | 86% | A, | —> | A■ | *Obviously,* | —> | *ǑBVIOUSLY*■ |
| | (48% | A | —> | A) | | | |
| DISAGREE | 14% | A, | -/-> | A■ | | | |
| | 52% | A | —> | A■ | *Obviously* | —> | *obviously*■ |
| Medial position (M) | | | | | | | |
| AGREE | 33% | ,A, | —> | ■A■ | *,obviously,* | —> | ■*ǑBVIOUSLY*■ |
| | (69% | A | —> | A) | | | |
| DISAGREE | 67% | ,A, | —> | A■/■A | *,obviously,* | —> | *obviously*■<br>■*obviously* |
| | 31% | A | —> | A■/■A | *obviously* | —> | *obviously*■<br>■*ǑBVIOUSLY*■ |
| Final position (E) | | | | | | | |
| AGREE | 17% | ,A | —> | ■A■ | *,obviously* | —> | ■*ǑBVIOUSLY*■ |
| | (69% | A | —> | A) | | | |
| DISAGREE | 83% | ,A | -/-> | ■A■ | | | |
| | 31% | A | —> | ■A■ | *obviously* | —> | ■*ǑBVIOUSLY*■ |

increased distance from initial position. Lack of agreement was manifested in two different ways: the adverbial comma in LOB was not matched by a tone unit boundary in LLC and, conversely, the adverbial was marked off by a tone unit boundary in LLC with no matching comma in LOB. Moreover, mid-clause adverbials in LOB were often provided with a comma on both sides, while only one tone unit boundary preceded or followed the adverbial in LLC (for the same result, see Altenberg 1987a:97-99).

Judging by these results, a general answer to the question whether an adverbial comma is a reliable cue to prosodic segmentation can be expressed as follows:

'Yes' if the adverbial is initial;
'Doubtful' if the adverbial is medial;
'No' if the adverbial is final.

**Figure 9:2.** *However* and *now* in initial position: correlation between comma and tone unit boundary per adverbial function.

| | LOB | LLC |
|---|---|---|
| ACcon | *However*, he couldn't make it | ‖*HOWEVER*■ he ‖couldn't ··· |
| AApro | *However* the tour was organized it became a success | ‖*HOWEVER* the tour was··· |
| ASint | *However* much he tried he couldn't get there | ‖*however* MUCH■ ··· |

| ITEM | TAG | LOB | % | LLC | % | EXAMPLE |
|---|---|---|---|---|---|---|
| *however* | ACcon | A, | 93% | A■ | 74% | HOWEVER■ |
| | AApro | A | 100% | A | 100% | however |
| | ASint | A | 100% | A | 100% | however |
| *now* | ACtra | A | 71% | A | 42% | now |
| | AAtim | A | 13% | A | 54% | NOW |

However, since some items can serve more than one function in the same position, and since different functions may be characterized by specific punctuation/prosodic conventions, a general answer like this is not sophisticated enough to be useful. Note also that the absence of a comma does not guarantee absence of a tone unit boundary, which means that commas in writing are never a sufficient guide to prosodic boundaries.

Let us consider *however* and *now* in sentence-initial position. Punctuation (a comma in LOB) and prosody (a tone unit boundary in LLC) distinguished the conjunct *however* from the process adjunct and intensifier adjunct *however*, as exemplified in Figure 9:2. In most cases in the written data, only contextual factors, but not comma, distinguished different functions of *now*. In the spoken material, on the other hand, prosodic features often served to mark off the time adjunct function from the transitional conjunct function:

| (3) | | LOB | LLC |
|---|---|---|---|
| | ACtra | *Now*(,) would you believe it? | ‖*now* would you BELIEVE it■ |
| | AAtim | *Now* she is in China. | ‖*NOW*(■ ) she is in ‖CHINA■ |

The tendency to punctuate the transitional conjunct *now* was fairly strong (42%). The tendency for a prosodic break to occur after the initial time adjunct *now* was stronger (46%), but maybe not strong enough for us to conclude that prosody marks the distinction. The presence or absence of a nuclear tone would be a safer criterion. Generally, the time adjunct *now* carries a tone while the transitional conjunct does not (cf intonation below).

Meyer (1986:47), following Greenbaum (1969), emphasizes that 'punctuation can be used to distinguish various types of homonymous adverbs, specifically adjuncts that can function also as conjuncts and disjuncts', as in

> He did it, *naturally*.
> He did it *naturally*.

and adds that 'one reason that punctuation did not regularly distinguish homonyms is that the homonymous forms do not occur in the same position', for example

> *Clearly* we could see it.
> We could see it *clearly*.

But this way of reasoning does not apply to *now*, which occurs initially both as a transitional conjunct and as a time adjunct.

In speech, not only tonicity (place of nuclear tone) but also tonality (type of tone) can parallel punctuation. Just as a prosodic break may distinguish one adverbial function from another (as with sentence-initial *now*), so type of tone can serve to distinguish the conjunct *still* from the time adjunct *still* (4):

(4) ACcon ‖STILL■ ‖no I ENJOY it■

AAtim ‖STILL■ we ‖have this deadlock SITUATION■

Both tonicity and tonality vary with the position of the adverbial in the sentence, and so does punctuation, but not necessarily in a one-to-one fashion.

The results of the examination of prosodic patterns per item, grouped according to adverbial subcategory, are presented in Figure 9:3 (the layout of which follows Allerton & Cruttenden 1976:41, 1978:161). Notice that the results do not always agree with the recommendations in Figure 9:5 which are partly based on Allerton & Cruttenden's results (for reasons already explained).

**Figure 9:3**. Prosodic patterns per category and item in the data.

| TAG | POSITION Initial | Medial | Final | ITEM |
|---|---|---|---|---|
| AApro | *0- | | ↘* | clearly, however |
| AAtim | *↘■, | _∨_ | ↘*,■ ∨* | again |
| | | | | eventually |
| | | | | now, still, yet |
| ACadd | *∨- | -∨- | | in particular |
| | | | | moreover |
| ACapp | *0-,*↘■ | -0- | tone* | especially |
| | | | | for example |
| | | | | for instance, thus |
| ACcon | *↘■,*0- | -0-,*↘■ | ↘* | after all |
| | | | | anyway |
| | | | | however, instead |
| | | | | nevertheless |
| | | | | yet |
| ACenu | | ↗_ | | by contrast |
| | | | | on the other hand |
| AClis | *↘■,*0- | -0- | ■↘* | also, finally |
| | | | | too |
| ACrep | *0- | -0- | -0,tone* | as a matter of fact |
| | | | | in fact |
| ACres | *↘■ | | | so |
| ACsum | *0- | -↘ | -0* | therefore |
| ACtra | *↘■,*0- | | | again, now |
| ADcnt | *∨■,*↘■ | -0- | tone*,-0* | apparently |
| | | | | basically |
| | | | | clearly, indeed |
| | | | | maybe |
| | | | | of course |
| | | | | perhaps |
| | | | | fortunately |
| | | | | unfortunately |
| ADsty | *∨■,*↘■ | -0- | ↗* | frankly |
| | | | | generally |
| | | | | honestly |
| | | | | presumably |
| AQpol | *↘■ | | -0* | please |
| AQres | *↘■ | | | indeed |
| AQsof | *∨■ | -0- | -0* | as it were |
| ASfoc | *0- | ■↘■ | ↘* | obviously |
| ASint | *0- | ↗■ | ↘* | indeed |

| | |
|---|---|
| * | indicates prosodic break 'by default', eg *PLEASE■ = separate TU |
| tone | stands for various types of tone, none of which predominates |
| 0- | indicates non-nuclear syllable |

Unfortunately, these results are too much a reflection of individual speech behaviour in particular speech situations to constitute a really helpful tool for writing automatic segmentation rules. The answer is not necessarily more data, but rather a different type of data. An alternative likely to give better results than impromptu conversation is non-interactive speech characterized by more regular prosodic breaks and a more neutral intonation, such as we may expect to find in the reading of radio news by professional broadcasters.

The following factors have been found to affect the use of the comma:

- breaks that would cause a tone unit boundary in speech
- emphasis
- length and complexity of the adverbials
- degree of clause integration
- position in the sentence

It is fairly obvious that there exists a certain parallelism between comma and tone unit boundaries (cf eg Altenberg 1987a:94ff). According to *A Manual of Style* (1969:111), 'commas should be used to set off interjections, transitional adverbs, and similar elements that effect a distinct break in the continuity of thought'. Such words would be set off by tone unit boundaries in speech. Meyer's study (1986:76) showed that adjuncts and disjuncts that rarely occupy a separate tone unit in speech were unpunctuated in 90% of the cases in the Brown corpus. At the same time, the relation between punctuation and prosody is both weak and unsystematic, as pointed out by Meyer (1986:69).

Adverbials that rarely occupy a separate tone unit are sometimes punctuated to match emphatic intonation in speech (cf Meyer 1986:76). The fact that long and complex adverbials are more likely to be punctuated than short and simple adverbials has been pointed out both by Quirk et al (1985:1626) and Meyer (1986:64), and that the same is true for the occurrence of tone unit boundaries is shown by Altenberg (1987a:92ff). In Meyer's material, short and simple adverbials were punctuated one third of the time. In Altenberg's material only complex adverbials in sentence-medial position were enclosed by tone unit boundaries; this agrees only partly with my results, which indicate that *in particular* was set off by a tone unit boundary on both sides but not *after all*, *by contrast* and *on the other hand*.

Closely integrated adverbials, such as adjuncts, were unpunctuated 54% of the time in Meyer's material, while the more loosely integrated disjuncts and conjuncts were unpunctuated only 36% of the time. Surprisingly, Altenberg's study of one of the LLC texts showed that adjuncts were nearly always separated by a tone unit boundary, while disjuncts tended to be prosodically integrated (1987a:95). In my study, disjuncts and generally conjuncts were separated by a tone unit boundary in initial position, while there was no clear tendency as regards adjuncts.

Position was related to complexity both in Meyer's material (1986:36) and Altenberg's material (1987a:94ff): the longer and more complex the adverbial, the stronger its tendency to be punctuated/separated by a tone unit boundary. In my material, this tendency was more pronounced for initial adverbials than for adverbials in other positions. Altenberg's conclusion with regard to initial adverbials and text-to-speech conversion that 'the punctuation of the input text can probably be used as a reasonably reliable segmentation cue' (1987a:95) only serves as a confirmation.

A further important factor for the occurrence of tone unit boundaries is speed of delivery (Altenberg 1987a:100); the slower the speech, the greater the chances of a tone unit boundary.

All these factors must be taken into account when we consider rules for automatic segmentation. But there is another factor that I should like to bring up again, the problem of homonymy. The fact that one item often serves more than one adverbial function, even in the same position and sometimes without distinctive markers, complicates the automatic tagging of the written text. It is therefore necessary to introduce a number of frequency-based generalizations in the tagging system. One such case is *now* which, as we have seen, can serve both as a time adjunct and as a conjunct in sentence-initial position; since it is more often a conjunct, the automatic word-class tagger treats it as a conjunct in the first place.

## 9.3 Rules for segmentation

Predictive rules, or rather recommendations, for prosodic segmentation can be calculated on the basis of the typical patterns per item, position and function (as suggested in Figure 9:4), if the tendencies are clear.

Figure 9:4 should be read as follows. *After all* served as a conjunct in all three positions. At the beginning of a sentence, it was generally followed by a comma in LOB and by a tone unit boundary in LLC, where it also constituted a separate tone unit. In mid-position, 56% had a comma on both sides but only one (succeeding) tone unit boundary. Neither comma nor tone unit boundary preceded final *after all*. *Clearly* was identified as a content disjunct at I and M

**Figure 9:4**. Rule calculation.

| ITEM | TAG | POSITION | COMMA | TU■ | RULE |
|---|---|---|---|---|---|
| after all | ACcon | I | 88% | 64% | after ↘ALL■ |
| | ACcon | M | 56% | 53% | after all■ |
| | ACcon | E | 0% | 29% | after ↘ALL |
| clearly | ADcnt | I | 14% | 29% | ■clearly |
| | ADcnt | M | 0% | 56% | ■clearly |
| | AApro | E | 0% | 70% | CL↘EARLY■ |

and as a process adjunct at E. It was occasionally punctuated in initial position, but never in medial and final position, and there was often a tone unit boundary (mostly preceding) in medial position.

## 9.4 Conclusion

When there was complete agreement between punctuation and prosody in the two corpora, rule prediction was easy, but there were numerous other cases where there was no such agreement. To what extent should a comma that is not matched by a tone unit boundary be taken to indicate prosodic separation? And what are we to do with cases where tone unit boundaries tend to occur in speech without a matching comma in writing? Furthermore, it must be kept in mind that, although there are a few, almost obligatory, punctuation rules and the comma is often used according to fairly strict conventions, this is not done consistently and not always in agreement with the segmentation of genuine talk. It would therefore be inappropriate to take punctuation practice alone as a criterion of a prosodic break. Nor can it be taken for granted that tone unit boundaries in conversation reflect the most adequate segmentation principles for text-to-speech conversion. Prosody, which is linked both to the speaker and to the speech situation, is considerably more flexible than punctuation and subject to all kinds of idiosyncratic behaviour.

Therefore, the rules that are recommended for the TESS lexicon in Figure 9:5 are not always the immediate result of the findings in the data, as displayed in Figure 9:3, but should rather be seen as a compromise, influenced not only by previous research but also by linguistic intuition and common sense.

**Figure 9:5**

| ITEM | FREQUENCY LOB | FREQUENCY LLC | CORRELATION I | CORRELATION M | CORRELATION E | RECOMMENDATION I | RECOMMENDATION M | RECOMMENDATION E | TAG |
|---|---|---|---|---|---|---|---|---|---|
| after all | 17 | 37 | + | ! | o | ↘■ | -0■ | ■↘ | ACcon |
| again | 663 | 403 | * | | | ↘■ | | | AAtim |
| again | 14 | 36 | + | | | ↘■ | | | ACcon |
| also | 944 | 238 | - | | | ↘■ | | | AClis |
| anyway | 58 | 280 | + | + | o | ↘■ | ■ ∨■ | ↗ | ACcon |
| apparently | 69 | 57 | o | | - | ∨■ | -0- | -0 | ADcnt |
| as a matter of fact | 8 | 15 | + | + | o | ∨■ | ↗■ | -0 | ACrep |
| as it were | 7 | 34 | | ! | o | ∨■ | -0- | -0 | AQsof |
| basically | 12 | 33 | | + | o | ∨■ | -0- | ↗ | ADcnt |
| clearly | 23 | 17 | o | | | ↘■ | -0- | | ADcnt |
| clearly | 101 | 17 | | * | o | | | ↘ | AApro |
| especially | 167 | 10 | o | | * | 0- | ■ ∨■ | ■ ∨ | ACapp |
| eventually | 71 | 24 | * | o | o | ∨■ | ∨ | ∨ | ACtra |
| finally | 166 | 28 | - | o | | ↘■ | -0- | ↗ | AClis |
| for example | 126 | 49 | * | ! | - | ↘■ | -0- | -0 | ACapp |
| for instance | 86 | 53 | + | ! | * | ↘■ | -0- | -0 | ACapp |
| fortunately | 34 | 16 | * | * | | ∨■ | ■ ∨■ | -0 | ADcnt |
| frankly | 7 | 20 | + | | | ∨■ | -0- | ↗ | ADsty |
| generally | 130 | 27 | * | o | | ∨■ | -0- | ↗ | ADsty |
| honestly | 1 | 15 | * | o | o | ↘■ | ↘ | ↘ | ADsty |
| honestly | 6 | 2 | | o | o | | ↘ | ↘ | AApro |
| honestly | | 2 | | | | ∨■ | | | AQres |
| however | 15 | | o | | | 0- | | | AApro |
| however | 511 | 36 | + | ! | o | ↗■ | ■ ↗■ | ■ ↗ | ACcon |
| however | 40 | 8 | o | | | 0- | | | ASint |
| in fact | 146 | 465 | - | | o | ↘■ | -0- | ↗ | ACrep |
| in particular | 58 | 14 | * | | | ∨ | ↘■ | ■ ∨ | ACadd |
| indeed | 162 | 104 | * | o | | ↘ | -0- | ↘ | ADcnt |
| indeed | 85 | 60 | | * | * | | ↗■ | ↘ | ASint |
| indeed | 2 | 7 | * | | | ↘■ | | | AQres |
| instead | 165 | 9 | * | | o | ↘■ | | ↘ | ACcon |
| maybe | 82 | 45 | * | * | | ∨■ | -0- | -0 | ADcnt |
| moreover | 55 | 3 | + | | | ∨■ | ■ ∨■ | | ACadd |

| ITEM | FREQUENCY | | CORRELATION | | | RECOMMENDATION | | | TAG |
|---|---|---|---|---|---|---|---|---|---|
| | LOB | LLC | I | M | E | I | M | E | |
| nevertheless | 92 | 22 | + | o | | ↘■ | -0- | -0 | ACcon |
| now | 1458 | 959 | \ | | | ↘ | | | AAtim |
| now | 31 | 452 | o | | | 0- | | | ACcon |
| obviously | 124 | 169 | o | * | | ∨■ | ■↘■ | ■∨ | ASfoc |
| of course | 146 | 680 | + | o | - | ↘■ | -0- | -0 | ADcnt |
| on the other hand | 78 | 36 | + | + | | ↘■ | ■↗■ | ↗ | ACenu |
| perhaps | 406 | 291 | * | o | - | ↘■ | -0- | -0 | ADcnt |
| please | 94 | 117 | + | | - | ↘■ | ↘ | ↗ | AQpol |
| presumably | 46 | 50 | | o | | ∨■ | -0- | ↗ | ADcnt |
| probably | 270 | 316 | * | o | o | ∨■ | -0- | -0 | ADcnt |
| so | 2413 | 1220 | + | | | ↗■ | | | ACres |
| still | 10 | 13 | + | | | ↘■ | | | ACcon |
| still | 833 | 14 | * | | | ↗■ | | | AAtim |
| surely | 126 | 43 | o | o | | 0- | ↘ | ■↘ | ADcnt |
| therefore | 296 | 64 | + | ! | | ∨■ | -0- | -0 | ACsum |
| thus | 257 | 3 | o | | | 0- | | | ACapp |
| too | 923 | 49 | | | * | | | ■↘ | AClis |
| unfortunately | 44 | 25 | + | * | | ∨■ | ■↗■ | -0 | ADcnt |
| yet | 441 | 43 | o | | | 0- | | | ACcon |

\+ = comma in LOB and TU boundary in LLC

! = comma on both sides in LOB but TU boundary on one side in LLC

o = no comma in LOB and no TU boundary in LLC

* = no comma in LOB but TU boundary in LLC

\- = comma in LOB but no TU boundary in LLC

# Note

This chapter is a revised version of Stenström 1988.

# 10

# Graphic English prosody

## *Jan Svartvik*

The ability to choose an appropriate prosody is a basic feature in the native speaker's competence. Children acquire this ability without apparent difficulty, at an early age before they can read and write, whereas the performance of many foreign learners who have an excellent command of other areas of language proficiency (such as vocabulary, grammar and style) is marred by their use of a non-native prosody. Prosody - which is here taken to include segmentation of speech into chunks of information (tone units), pitch, loudness, stress and rhythm - is one of the least understood aspects of language description with many areas that have in fact been characterized as 'near-virgin territory' (Cruttenden 1986:xi).

The generation of artificial speech provides one means of finding out more about the properties of prosody. The advent of voice synthesizers has made it possible also for non-phoneticians to produce artificial speech. As yet, however, the effect of the droning voices of these fantastic speaking machines ranges from slight irritation to total incomprehension. There is a wide range of performance variation among text-to-speech systems, from Votrax Type-'N'-Talk with only 67.2 per cent correct identification to MRT Natural Speech with 99.4 per cent correct identification (Pisoni, Nusbaum & Greene 1985:1667). However, for all their limited performance, even the currently available machines can be used for learning about the prosodic parameters that affect artificially produced output. It seems reasonable to expect this kind of information to be useful not only for improving synthetic speech but also for understanding the nature of natural, human prosodic features, which must surely be the yardstick by which the quality of synthetic speech has to be measured.

In our studies of spoken discourse within the TESS project, we wanted to have, in addition to synthetic voice output, a graphic representation of some of the prosodic features that are recognized in the auditory prosodic transcription in the London-Lund Corpus. This paper describes one such experiment.

Five students of Computer Science at Lund Institute of Technology, under the supervision of Ingemar Dahlstrand, were invited to participate in a subproject called 'Graphic Prosody'. This task involved writing programs that would automatically provide information about tone units as analyzed in the Corpus. The programs produce two types of graphic prosodic output: one with the letters of the text arranged in the squiggly mode à la Bolinger, the other with the prosody indicated as a four-level graph curve. In addition, they produce some basic statistical information for each tone unit (number of characters, words, and syllables). The programs were written in Turbo Pascal for the IBM AT PC with the Enhanced Color Display.

The input to the program can be described as consisting of the following linguistic levels:

| | |
|---|---|
| TEXT: | the orthographic representation of the spoken material as recorded on audiotape. Needless to say there is no use made of conventional punctuation as used in writing (periods, commas, etc). |
| PROSODY: | the auditory analysis of the prosody used in the recordings (see pp 12ff). |
| WORD CLASS: | the word class tags given to the words in the texts (see pp 87ff). |
| PHRASE: | the analysis of the texts in terms of five types of grammatical phrase: adverb phrase, adjective phrase, verb phrase, noun phrase, and prepositional phrase (see p 95). |

This information was input from two files. Figure 10:1 shows the beginning (tone units 1-16) of file S1206.BAL, which is text S12.06 at text level. This includes representations of orthographic words (*well*, *rather than*, etc); prosodic analysis, ie onset, single stress, booster, nucleus, etc; and word class tags, eg 'DW' for discourse item, 'CE22' and 'CE21' for complex conjunction, 'VA+0' for base form of verb, 'TF' for determiner, etc (see pp 96ff). Since this file has all three levels mixed it is rather difficult to read for the human eye.

**Figure 10:1.** File S1206.BAL

```
1206000001 a     well<DW> ^rather<CE22> than<CE21> 'give<VA+0> a<TF> :t\alk<NC>
                 a'bout#<PA>
1206000002 a     the<TA> ^history<NC> of<PA> Stoke<NP22> :P\oges#<NP21> .
1206000003 a     I<RA> ^felt<VA+D> it<RC> 'might<VM+9> be<VB+0> a<AQ22>
                 :little<AQ21> more<AF> :\/interesting<JA> {to<PA> you<RC>
                 ^/all#}#<RP> -
1206000004 a     to<PD> ^hear<VA+0> a'bout<PA> my<TB> :own<JM> l\ife#<NC> -
1206000005 a     ^lived<VA+D> and<CA> :growing<VA+G> \up#<AP>
1206000006 a     in<PA> ^this<TD> . 'wonderful<JA> !village<NC> of<PA>
                 Stoke<NP22> :P\oges#<NP21> - - .
1206000007 a     I<RA> at^t\/ended#<VA+D>
1206000008 a     ^Stoke<NP22> Sch\ool#<NP21> .
1206000009 a     and<CA> I<RA> ^must<VM+8> s/ay#<VA+0>
1206000010 a     I<RA> was<VB+5> ^t\aught#<VA+N>
1206000011 a     ^very<AI> th/oroughly#<AW>
1206000012 a     the<TA> ^three<JR> /Rs#<NX+2> -
1206000013 a     ^funnily<AW> e'nough<AE> my<TB> :f\ather#<NC> .
1206000014 a     ^w\ent<VA+D> to<PA> the<TA> 'same<JM> sch/ool#<NC> .
1206000015 a     and<CA> ^he<RA> was<VB+5> 'one<JR> of<PA> the<TA> 'first<JQ>
                 :p\upils#<NC+2> -
1206000016 a     be^fore<PA> th\at#<RD> .
```

**Figure 10:2.** File S1206.PHR

```
1206000001 a
WORD    well rather than give       a  talk       about
TAG     DW   CE2         VA+0       TF NC         PA
PHRASE                   VPH1:base  NPH:det,nhead

1206000002 a
WORD    the history of Stoke Poges
TAG     TA  NC      PA NP2
PHRASE  NPH:det,nhead,prepmod

1206000003 a
WORD    I       felt      it      might be        a little more interesting to  you all
TAG     RA      VA+D      RC      VM+9  VB+0      AQ2      AF   JA          PA  RC  RP
PHRASE  NPH:per VPH2:past NPH:per VPH1:past,modal JPH                       PPH

1206000004 a
WORD    to hear  about my own life
TAG     PD VA+0  PA    TB JM  NC
PHRASE  VPH1:inf PPH
```

This should not worry us too much, however, since this representation is intended solely for the computer.

Figure 10:2 shows the beginning of file S1206.PHR, including tone units 1-4 and now arranged in a more user-friendly fashion than in Figure 10:1 with one tone unit of the text (WORD) per line and the word class analysis (TAG) and grammatical phrase analysis (PHRASE) appearing below the text. Thus, at the word class level, *I* is a personal pronoun in the subjective case (RA) which, at the grammatical phrase level, forms a noun phrase consisting of a personal pronoun (NPH:per); at the word class level, *felt* is a main verb in the past tense (VA+D) which, at the grammatical phrase level, forms the verb phrase in the tone unit, and the phrase is simple and in the past tense (VPH2:past); *might be* is the first verb phrase (counting from right to left in the tone unit) and analysed as VPH:past, modal.

Since the original documentation of the program Prosody by Isaksson et al (1986) is unpublished and written in Swedish, it will be convenient to give a brief presentation of it here. The program opens up with the main menu:

*File name* ........
*Output device* ....
*Display mode* .....
*Statistics* .......
*Exit* ............

*File name* asks for the name of the file to be processed; *Output device* offers three possibilities in a submenu:

*Screen* ..........
*Printer* .........
*Votrax* ..........

*Screen* gives the analysis on the screen; *Printer* provides hard copy output; *Votrax* activates the speech synthesizer, so that the analysed tone unit can be heard as well as seen. If *Printer* and/or *Votrax* are selected there is a further choice available: the output can be continuous, ie writing or speaking one tone unit after the other without interruption, or non-continuous, if a tone-unit-by-tone-unit analysis is preferred. When the former choice is made, the number of desired tone units can be stated, say the first 75 tone units in a text.

Going back to the main menu, the option *Display mode* offers the following possibilities:

*Text line*, ie the text is presented in simple orthographic transcription;
*Tag line*, ie word class level;
*Phrase line*, ie grammatical phrase level;

*Text prosody line*, ie a simple prosodic representation by means of a four-level arrangement of the letters of the words in the tone unit;
*Graphic line*, ie a prosodic representation by means of a graphic display;
*Statistics*, ie access to some basic statistical information about the processed tone units.

If the last option is selected, the submenu for Statistics includes the following choices:

*Number of tone units* .......
*Number of words* ...........
*Number of syllables* ........
*Number of characters* .......

When leaving the program by opting for *Exit* in the main menu, the program asks if you want to save default values, in which case the selected output format is saved for the next time the program is invoked.

The main reason for initiating the subproject Prosody was to test out different ways of producing graphic representation of prosodic features. The chief instructions given to the students of computer science were the following:

Within each tone unit,

- set tone unit onset at level 2;
- lower each successive syllable one level unless marked for booster, stress or nucleus; in which case
- raise syllables marked with booster (:) or single stress (') one level;
- raise syllables marked with high booster (!) or double stress (") two levels;
- raise syllables marked with extra high booster (!!) three levels;
- indicate nuclear pitch change (rising and falling tone) in some appropriate way; and
- use four levels (1, 2, 3, 4); if raising according to the above instructions exceeds the given range, the top level (4) is to be used.

Compared with the output printed here, the screen version of Prosody looks far more attractive with different colours helping to distinguish the various types of analysis. Figure 10:3 shows some examples of output on paper (tone units 1-4). In addition to the text, word class and grammatical phrase levels from the input, the figure gives for each tone unit, both a graphic contour (GRAPH) and a squiggly type of intonation display (PROS). Below the graph we can see the original prosodic notation as on the computer tape version of the London-Lund Corpus. Finally there is provided statistical information about number of characters, words and syllables.

**Figure 10:3.**

```
TEXT   well rather than give a talk about
TAG    DW   CE2         VA+0 TF     PA
                             NC
PHR                     VPH1:base
                             NPH:det,nhead
PROS         rat
                her
       well         than give   t
                             a  ALK about
GRAPH
```

```
ORIG   well ^rather than 'give a :t\alk a'bout#
Number of characters: 28     Number of syllables:  9
Number of words:       7     Number of toneunits:  1
```

```
TEXT   the history of Stoke Poges
TAG    TA  NC        PA NP2
PHR    NPH:det,nhead,prepmod
PROS       his
              to
       the      ry          P
                  of Stoke  OGES
GRAPH
```

```
ORIG   the ^history of Stoke :P\oges# .
Number of characters: 22     Number of syllables:  8
Number of words:       5     Number of toneunits:  1
```

```
TEXT   I felt it might be a little more interesting to you all
TAG    RA     RC VM+9  VB+0       AF   JA           PA RC   RP
         VA+D             AQ2
PHR    NPH:per   VPH1:past,modal                    PPII
         VPH2:past        JPH
              NPH:per
PROS     felt                                               ALL
              it might                         TING to
       I                be   little         RES         you
                           a        more INTE
GRAPH
```

```
ORIG   I ^felt it 'might be a :little more :\/interesting {to you ^/all#}# -
Number of characters: 44     Number of syllables: 15
Number of words:      12     Number of toneunits:  2
```

```
TEXT   to hear about my own life
TAG    PD VA+0 PA    TB JM  NC
PHR    VPH1:inf
               PPH
PROS      hear
               about    own
       to            my     l
                             IFE
GRAPH
```

```
ORIG   to ^hear a'bout my :own l\ife# -
Number of characters: 20     Number of syllables:  7
Number of words:       6     Number of toneunits:  1
```

The Prosody program is based on the tone unit, since we consider this prosodic chunk to be important in speech production and comprehension; hence it should be valuable to know more about the relation between its grammatical and prosodic properties. The statistical information coupled with grammatical and prosodic information is therefore an additional asset. Except for a few tricky cases (such as abbreviations), the program is quite successful in counting syllables. In any case, it should be possible to achieve at least a reasonable approximation with this method. So far, most counts of tone unit length have used number of words, which is probably less informative than number of syllables. In particular, the combination of these different ways of counting should be useful when combined with timing. In a study of ten texts in our corpus, it was found that the average tone unit length was 1.9 seconds and 4.5 words, and that

> speakers divide their utterances into chunks that vary surprisingly little from one communicative situation to another... Of the examined features [speech rate, seconds per tone unit, and words per tone unit], tone unit length is clearly the least variable feature in the material, while speech rate, for example, is almost three times as variable (Altenberg 1987a:25).

As a result of the simple instructions to the programmers, the two types of prosodic notation provided by the Prosody program are of course a great oversimplification and not in agreement with the Crystal model, as they now stand. There are three reasons why this system was still chosen. First, in view of what might appear to be a rather intimidating task to nonlinguists, it seemed better to ask the students specializing in computer science to put their main effort into devising the program framework rather than testing out different types of conversion of prosodic input to graphic output; second, in view of what was said at the outset of this chapter, it is not quite clear what is actually the relationship between prosodic features (for example between onset, booster, and stress); third, our Votrax speech synthesizer is a rather crude instrument (as indicated by the low figure for correct identification, p 267) and, considering the limited time available on the TESS project, we had in any case to concentrate our efforts on predicting reasonable chunking rather than intonation. What was a realistic aim to achieve was far from the 'definitive' program but one kind of research tool to be used for experiments with different types of representation of the original information contained in the auditory prosodic analysis of the London-Lund Corpus.

From the point of view of linguistic methodology, it seems worth stressing the possibilities that computers offer in churning out, obediently and tirelessly, whatever we ask them to do by specific instructions, thus providing us with a means to apply our analysis to large amounts of data, to identify relationships we might not otherwise discover, to check and recheck how well different analyses work - but also highlighting, in glaring detail on the colour screen, all the bugs in our design. All the steps in the analysis by computer reported here - for word class, grammatical phrase, and prosody on screen, paper and via synthesis - are fully automatic. While the machine can never equal man in terms of sensitive linguistic analysis, it can be very helpful in giving us certain insights which are easily missed in manual analysis through its outstanding characteristics: a capacity for endless repetition, ruthless consistency, and strict specification.

## Note

This chapter is based on Svartvik 1987. I am grateful to I. Dahlstrand and his students L. Isaksson, A. Landin, O. Persson, K. Tsilkos, and M. Wilson at the Department of Computer Science, Lund Institute of Technology, for their help in producing the Prosody program.

# 11

# Predicting text segmentation into tone units

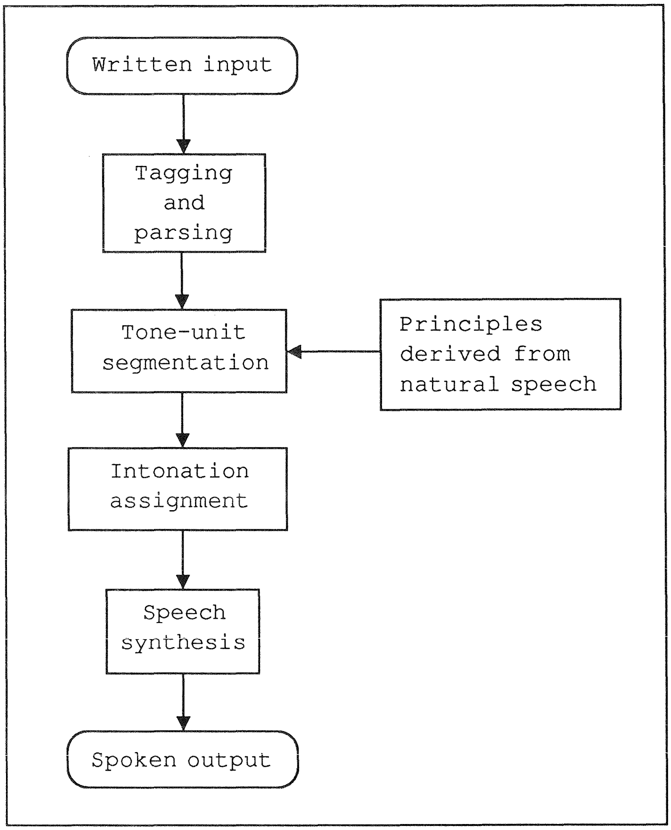
*Bengt Altenberg*

## 11.1 Introduction

A central task for the TESS project was to set up a system of rules that will automatically 'chunk' a written input text into prosodic segments resembling the information units or tone units a speaker produces in natural speech. Such a rule system will be presented in Chapter 12, but before we turn to these rules, it is necessary to examine the principles governing natural speech segmentation and see to what extent these principles can be reduced to a set of predictive, grammatically defined segmentation rules that can be incorporated in an automatic text-to-speech system. I will here give a brief report of a larger study carried out to determine the principles of speech segmentation in a prepared and partly scripted monologue (S.12.6) from the London-Lund Corpus (see Altenberg 1987a: Chapter 4). My intention is not to give a detailed account of the study, but rather to give a general idea of the approach that was used, the problems involved, and some of the results.

A simplified picture of the sequence of operations required in a text-to-speech conversion program is given in Figure 11:1. The written input text is first tagged and parsed automatically, ie analysed in terms of grammatical

**Figure 11:1.** Flowchart showing the principal components of text-to-speech conversation.




word class and syntactic structure (see pp 87ff). The output of the parser then serves as input to the segmentation rules, which chunk the grammatically analysed text into tone units on the basis of principles derived from natural speech. The tone units identified in this way mark the domain of the intonation rules in the following component. The TESS project was mainly concerned with the grammatical aspects of this process. In the study reported here the focus of interest is on the relationship between the first three components - the parser, the automatic segmentation rules, and the natural spoken data.

There are several problems connected with this relationship, and I will briefly touch on three of these. The first concerns the delicacy of the grammatical analysis produced by the parser. I have assumed a fairly advanced parser that is capable of assigning not only word-class and phrase-category labels to the input string, but also clause element functions (subject, verb, object/complement, adverbial) and certain clause functions (relative, nominal, adverbial, etc). It must be emphasized, however, that in an optimal system the 'depth' of the grammatical analysis will have to be determined by several factors, practical as well as theoretical. One such factor is the model of speech segmentation and intonation assignment adopted in the subsequent components of the system and the kind of cues that the rules of these components need to produce an acceptable output. In other words, the demands on the parser will partly depend on the outcome of studies like the present one. I will return to this problem later (p 286).

The second problem concerns the fact that segmentation principles derived from natural speech are applied to a written text. Speech and writing are obviously very different both in terms of processing and structure. We have tried to reduce this difference by using a prepared and partly scripted monologue as our preliminary model, but it is evident that even a prepared speech is primarily intended to be heard, and will therefore differ considerably from a text that is intended to be read.

The third problem concerns the difficulty of turning principles of natural speech segmentation, which is basically a semantic process, into predictive rules that must rely solely on the lexico-grammatical information produced by the parser. We may assume with Chafe (1987) that when a speaker is speaking, he successively 'activates' ideas in his mind; each of these ideas is encoded and brought into focus as 'new' information in a separate information unit and marked prosodically with a salient pitch movement (the nucleus). Hence each information unit can be said to be organized around a focal element. What the speaker chooses to bring into focus is primarily a semantic decision, though his way of presenting it is influenced by situational factors - cf Halliday's (1985b:278) distinction between a speaker-oriented thematic structure and a listener-oriented information structure, both of which are speaker-selected. This focusing process is very different from what a computer can be taught to do. A computer cannot interpret the input text semantically and chunk it accordingly (as a human reader would do); it can 'only' analyse the text grammatically and base its segmentation on rules referring to formally identifiable properties in the parsed text.

The problem is thus how the semantically based principles of natural speech segmentation can be reduced to grammatically defined cues that the computer can understand. Fortunately, the segmentation of natural speech also reflects grammatical principles. A speaker does not divide his utterances just anywhere, but tends to make his prosodic divisions at major syntactic boundaries, eg between clauses and major clause constituents (see eg Quirk et al

1964, 1985:1357ff, Brown 1977:84ff, Crystal 1975:15-22). The task is thus to examine the degree of correlation existing between prosodic and grammatical boundaries in natural speech and determine to what extent this correlation can be converted into predictive segmentation rules.

The approach to speech segmentation used in the present study is based on a predictive model devised by Crystal (1975:15-22), outlined here in Figure 11:2. Although this model was originally designed for conversational data and not intended for text-to-speech conversion, it makes use of a prosodic and grammatical framework that is well suited for the present material. The model operates cyclically in top-down fashion, scanning a grammatically analysed input string for possible prosodic 'breaking points' at three levels: sentence level (between clauses), clause level (between clause elements) and phrase level (between phrase constituents). At each level a distinction is made between 'basic' and 'expanded' structures. Basic structures are processed as a single tone unit (unless they are expanded at a lower level), while expanded

**Figure 11:2.** Crystal's (1975:16) model for assigning tone-unit structure to sentences.

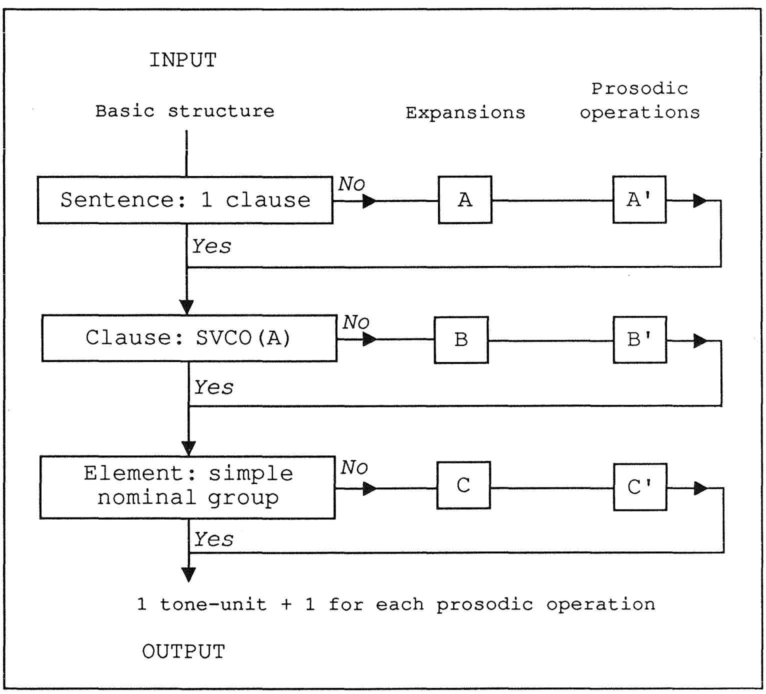


**Table 11:1.** Clause boundaries cooccurring with tone-unit boundary.

| TYPE OF CLAUSE BOUNDARY | TOTAL | TU BOUNDARY N | % |
|---|---|---|---|
| After initial clauses | 29 | 29 | 100 |
| Around medial clauses | 15 | 15 | 100 |
| Before finite adverbial clauses | 46 | 46 | 100 |
| Before adverbial *ing*-clauses | 14 | 14 | 100 |
| Before nonrestr. relative clauses | 26 | 26 | 100 |
| Before asynd. clause coordination | 15 | 15 | 100 |
| After postmodifying clauses | 67 | 66 | 99 |
| Before synd. clause coordination | 153 | 150 | 98 |
| Before nonfin. postmodifying clauses | 25 | 19 | 76 |
| Before restrictive relative clauses | 26 | 18 | 69 |
| After comment clauses | 13 | 9 | 69 |
| Before adverbial infinitive clauses | 12 | 8 | 67 |
| Before comment clauses | 13 | 8 | 62 |
| Before nominal *that*-clauses | 32 | 19 | 59 |
| Before nominal rel./interrog. clauses | 16 | 7 | 44 |
| Before nonfinite nominal clauses | 21 | 7 | 33 |

structures are subject to prosodic operations which assign a tone-unit boundary at the point of expansion according to rules that are specified for each level.

## 11.2 Tone-unit boundaries between clauses

At sentence level, prosodic boundaries may occur between coordinate clauses and between subordinate clauses and their matrix clause. It is necessary to make a number of distinctions here, since the likelihood that a tone-unit boundary will coincide with a clause boundary is determined by the clause types involved, their degree of cohesion, relative position, syntactic function, etc. I have generally used the distinctions made in Quirk et al (1985:Chapters 12-16).

The most frequent types of clause boundaries in the examined monologue are shown in Table 11:1. The clause boundaries are arranged according to their tendency to be accompanied by a tone-unit boundary, with those having high cooccurrence at the top of the table and those having low cooccurrence at the bottom. The table describes a scale of cooccurrence ranging from 100% to 33%. The scale is characteristically 'top-oriented', ie there is no type of clause boundary that consistently 'repels' a tone-unit boundary. This indicates that clause boundaries tend to constitute major breaks in the semantic and prosodic organization of speech and, we may assume, in the speaker's internal processing (cf Stenström's account of the distribution of pauses, p 236).

A large number of clause types show a consistent tendency to be set off by a tone-unit boundary. These regular cases, which comprise 67% of all clause boundaries in the text, include all initial and medial clauses, all coordinate clauses, and most 'peripheral' or loosely attached clause types. Most of the boundaries involved can be defined formally in one way or another, and can thus be incorporated in a predictive rule system. There are some exceptions, however, notably nonrestrictive relative clauses, which cannot be automatically distinguished from their restrictive counterparts. (Punctuation may be of some help in this respect, but is generally too inconsistent to be a reliable guide to prosodic segmentation. On the whole, the correspondence between punctuation and intonation is little investigated; for some observations, see pp 253ff, Cruttenden 1986:181f, Meyer 1986:92-94.)

The prosodically variable clause boundaries further down the scale (below the broken line) generally involve clause types that are more cohesive, and therefore less easily separable, than those at the top of the scale. If we contrast the two halves of the table, we can detect some grammatical reflexes of this greater cohesiveness. Apart from the restrictive relative clauses, we find in the bottom half a greater concentration of final nonfinite clauses, including all infinitive clauses. These are frequently reduced by ellipsis and therefore tend to be more cohesive than finite clauses. There is also a concentration of final nominal clauses, most of which have an object or complement function and consequently tend to be more closely attached to their matrix clause than adverbial clauses.

However, these are general tendencies only. Properties like 'restrictive', 'nonfinite', 'infinitive' or 'object function' have no reliable predictive value and are easily overruled by other factors (eg the status of the clause as a complex-transitive, pseudo-cleft or extraposed construction). The most important factor is not grammatical at all, but related to the information structure and the location of the information focus. If the speaker chooses to put his focus on a rhematic element in the matrix clause, the following subordinate clause is normally set off in a separate tone unit (with a focus of its own), but if he moves the focus forward to an element in the subordinate clause, there is generally no need for a prosodic separation: the whole sequence can be produced as a single information unit. This focusing principle, which applies to all final clause types, is illustrated in the following clause pairs containing nominal *that*-clauses (1) and postmodifying clauses (2):

(1) (a) and ‖I should ▵THINK■ the • ‖Adams ▵BROTHERS■ DE‖SIGNED it■
(S.12.6:789-791)

(b) I ‖think it was 'through 'her 'INSPIRATION■ that ‖POSSIBLY■ • the ‖Women's INSTITUTE■ (...) ‖really DE▵VELOPED■ (S.12.6:995-999)

(2) (a) ‖and it was the Δfirst TIME■ I ‖ever SAW■ the ‖big • Δstone JARS■ with ‖MUSTARD PICKLE■ (S.12.6:546-549)

(b) ‖AND■ • you ‖KNOW■ it's the ‖first 'time I've had ▵dealings with A▵MERICA■
(S.12.6:807-809)

In the (a) examples an element in the matrix clause is in focus (*think* and *time*, respectively) and followed by a tone-unit boundary; in the (b) examples the focus is postponed to the following subordinate clause and no separation is needed.

There are no doubt many additional reasons for this difference in focus placement, for example the need to highlight a particular element (as seems to be the case with *think* in (1a)), the speaker's fluency, speed of delivery, etc. These are all extralinguistic factors and not predictable as such, but it is interesting to note that the location of the focus generally has a grammatical correlate that may be used for predictive purposes. One such cue is the syntactic structure of the matrix clause. The likelihood that the speaker will place the focus on a matrix element obviously increases if the matrix clause contains a postverbal element - an object, complement or adverbial before the subordinate clause. Such postverbal elements usually convey new information and therefore easily attract focus:

(3) ‖and of Δcourse they'd ▹made Δvery "SURE■ that ‖no ENCROACHMENTS {‖HAPPENED■ }■ (S.12.6:893-894)

On the basis of this tendency it is possible to set up a 'Matrix Rule' which assigns a tone-unit boundary after matrix clauses having a SVC, SVO or SVA structure. This rule is not infallible, but if supplemented with a 'length factor' which prevents a division after matrix clauses of less than five words, its predictive accuracy is surprisingly high. This length factor may seem arbitrary, but it has clear statistical support in the text. It seems to reflect the fact that, if the speaker has not anticipated the end of his utterance by the time he has produced his first four words (which is the average length of tone units in this text), he is forced to make a prosodic break very soon.

The Matrix Rule and the length factor are good examples of how essentially nongrammatical principles of speech segmentation can be translated into formally defined properties that can be used for predictive purposes. Together with certain other grammatical cues (that will not be described here) this rule increases the predictability of the 'variable' clause boundaries in the lower half of Table 11:1 considerably. If we include the invariably separated clause types (the upper half of Table 11:1), 95% of all clause boundaries in the text can be covered by formally defined rules that accurately predict the presence or absence of a tone-unit boundary in 93% of the cases.

## 11.3 Prosodic separation of clause elements

After the sentence cycle, the segmentation rules move on to the clause level and check the input string for possible complexity in clause-element structure. Here Crystal's model makes a distinction between 'basic' and 'expanded' clauses. The former are defined as consisting 'maximally of the elements Subject + Verb + Complement and/or Object, with one optional Adverb, in this order' (1975:16). If we conflate objects and complements under one symbol C, we can represent this basic pattern roughly as

S V (C) (C) (A)

where the bracketed symbols are alternative (optional or obligatory) clause elements. Expanded clauses involve various complications of this structure in the form of additions or reordering. The most frequent type of expansion in the present text is the appearance of an extra adverbial, which may occur initially, medially (in various positions) and finally (after the single adverbial allowed in basic clauses).

Crystal's model only predicts a prosodic division of expanded clauses under certain conditions. In the present text, expanded clauses are certainly more frequently divided (82%) than basic clauses (47%), but the situation is much less clear-cut than Crystal's model presupposes. In basic clauses, elements are quite often prosodically separated, especially if they occur in thematic or extreme rhematic position (final adverbials 30%, subjects 19% and final objects/complements 16%), whereas penultimate objects/complements are rarely separated from the verb (9%). Separation is normally blocked if the element is a pronoun, but otherwise it is difficult to explain the segmentation in grammatical terms. What determines the segmentation is always the information structure of the clause and the location of the focus. Subjects, for example, which are brought into focus when they convey new information (eg when they signal a new topic), are always separated prosodically if the rheme also contains focal information:

(4) my ||MÒTHER■ • was a ||very INDÙSTRIOUS■ • ||WŎMAN■ (S.12.6:71-73)

(5) ||my MÒTHER was a 'great • 'Women's ÍNSTITUTE■ (S.12.6:1012)

In these examples *my mother* signals a new (sub)topic in both cases, but while the complement is new in (4) it is given in (5). Hence there are two new elements in (4) but only one in (5), and only the former needs to be segmented into two information units.

Similar conditions determine the treatment of a final adverbial, which is separated if it conveys new information and there is a focal element earlier in the clause (6), but not otherwise (7):

(6) I ||happened to 'spend a ▵[sko] 'a ▵HÒLIDAY■ in ||SHÈTLAND■ (S.12.6:652-653)

(7) this ||made no △DÌFFERENCE • to this GÍRL■ (S.12.6:473)

However, a prosodic separation of basic clause elements never seems to be absolutely necessary. Many of the separated cases can be ascribed to hesitation or slow delivery and, in the great majority of the examples, a removal of the tone-unit boundary produces an improved 'reading'. This is illustrated in (4) above, where the tone-unit boundary after *mother* can be replaced by a high onset (to signal the new topic), and in (6), where a single nucleus on *Shetland* seems to be quite sufficient.

By contrast, adverbials expanding the basic clause pattern generally have a much stronger tendency to be prosodically separated, especially in initial (73%) and final (51%) position. The possibility of predicting the segmentation of these adverbials by means of neatly formalized rules is complicated by their great formal and functional diversity. Short medial adverbials (mainly adjuncts or subjuncts like *always*, *never*, *certainly*, *also*) are seldom prosodically separated (except when emphasized), whereas adverbials realized by prepositional phrases usually are. Initial adverbials, which generally have a grounding, connective or attitudinal function (as adjuncts, conjuncts or disjuncts), are normally set off in a separate tone unit if they are polysyllabic, and final adjuncts (following another adverbial) are usually separated if they convey new information and have a certain 'weight'. The simplest treatment of all these variants seems to be to disregard their function and predict their prosodic behaviour in terms of position and form. A predictive rule that separates initial adverbials 'heavier' than a closed-class adverb, and medial and final adverbials realized by a prepositional phrase containing at least one open-class word, is fairly successful, but there is no doubt that this rule can be sharpened further.

## 11.4 Prosodic division of phrases

At the phrase level, Crystal makes a distinction between 'simple' phrases consisting of a head maximally modified by 'one adjectival premodifier and/or one prepositional phrase postmodifier', and 'multiple' phrases realized by more complex structures (1975:16f). The latter type thus includes all coordinate and appositive phrases, all heavily premodified phrases, and all phrases containing more than one postmodifier. Simple phrases are not prosodically divided, whereas multiple phrases are set off prosodically (but not divided) under certain conditions.

In the present text, the tendency to split phrases prosodically is generally greater than Crystal's model predicts. This tendency applies to both simple and multiple phrases, but is especially noticeable with simple postmodified phrases, particularly when both the head and postmodifier convey new information. However, the separation of these cases is generally optional and due to slow or hesitant speech.

Premodified phrases - simple as well as multiple - are also split occasionally (generally before the noun phrase head). These divisions are grammatically unpredictable, but since they are mainly due to hesitation or special emphasis, I have disregarded them as 'negligible exceptions'.

The multiple phrases that are easiest to predict prosodically are those involving coordination and apposition. Jointly modified syndetic coordinations (roughly defined as those lacking a premodifier before the second conjoin) are rarely divided (*the great cries and encouragements*), whereas separately modified syndetic coordinations

(8) he ||had • ∆white HA↗IR■ and ||this black ▵FA↘CE■ (S.12.6:978-979)

and asyndetic coordinations

(9) the ||CO↘WSHEDS■ the ||MA↘NGERS■ the di||viding ∆WA↘LLS■ (S.12.6:597-599)

always are. Non-restrictive appositions are also consistently divided

(10) that "‖LÒVELY MÁNSION■ · the ‖Manor HÒUSE■ (S.12.6:1034-1035)

whereas the few restrictive appositions that occur in the text are generally not (*we children*).

## 11.5 Conclusion

The approach to automatic speech segmentation that has been outlined here involves two successive tasks. The first task is to discover principles of natural speech segmentation that are regular enough to be used for predictive purposes. The second task is to convert these principles into formally defined predictive rules.

Each of these tasks presents problems. Failure to discover patterns of segmentation in the data inevitably leaves gaps in the *coverage* of the predictive rules. Reliance on weak or irregular patterns, or failure to formalize the rules properly, reduces the *success rate* of the rules. In Table 11:2, I have tried to evaluate the results of the study in these two terms. As the table shows, the coverage of the suggested rules is quite satisfactory for boundaries between clauses and clause elements, but less satisfactory for the segmentation of multiple phrases. The problematic categories here are the pre- and postmodified phrases, which are often divided without a discernible pattern.

The rules are most successful in handling the prosodic separation of clauses, but less successful in predicting the separation of phrases and, especially, clause elements. The reason for this difference is that the prosodic

**Table 11:2.** Coverage and success rate of proposed segmentation rules.

| TYPE OF BOUNDARY | COVERAGE | SUCCESS RATE |
|---|---|---|
| 1 Between clauses | 95% | 93% |
| 2 Between clause elements | | |
| a) Basic clauses | No segmentation predicted | |
| b) Expanded clauses | 94% | 70% |
| 3 Between phrase constituents | | |
| a) Simple phrases | No segmentation predicted | |
| b) Multiple phrases | 78% | 80% |

breaks at clause boundaries can generally be captured in grammatical terms, whereas the segmentation of phrases and clause elements is less regular and often easier to define in textlinguistic or discourse-functional terms than in strictly grammatical terms. In addition, the segmentation of phrases and clause elements is more exposed to pragmatic and extralinguistic influences, such as emotive highlighting, fluency, and speed of delivery. Some of these factors give rise to truly unpredictable tone-unit boundaries (eg emotive highlighting), others create what I have called 'negligible exceptions' to otherwise regular patterns. The latter I have ignored in the hope that they will disappear in faster and more fluent speech.

Table 11:2 also indicates where further research is most needed. The prepared monologue that was used for the present study obviously gives a very limited picture of natural speech segmentation. However, the results provide a good starting-point for further investigations, both generally and in the areas of poor rule coverage and success rate noted above. Truly unpredictable cases will remain, but as more extensive material is examined, there is no doubt that the rule system can be enriched considerably and improved both in terms of reliability and validity.

Finally, I will return to one of the problems mentioned in the introduction to this chapter. The approach that has been tested here has been a top-down cyclical model that presupposes a high-level grammatical analysis of the input text. This makes heavy - and perhaps unrealistic - demands on the efficiency of the parser. It would therefore be worthwhile to explore alternative models using a shallower grammatical analysis (primarily restricted to the word and phrase levels - cf pp 289ff) or even a bottom-up approach of the type investigated by Knowles & Lawrence (1987). Progress is possible only if we are prepared to try different possibilities and evaluate the relative efficiency of these. Our knowledge of speech production and processing is still very limited, but the arrival of computers and speech synthesizers has provided us with new and fruitful tools for testing ideas and hypotheses derived from natural data. As Cruttenden (1986:184) has put it, it "is an exciting time for intonationists".

## Note

This chapter is a slightly revised version of a paper (Altenberg 1987b) presented at the Seventh International Conference on English Language Research on Computerized Corpora (ICAME 7), Amsterdam, 8-11 June 1986.

# 12

# Automatic text segmentation into tone units

*Bengt Altenberg*

## 12.1 Introduction

An automatic text-to-speech conversion program that aims at producing natural-sounding speech must contain, as one of its components, a set of rules that 'chunk' the input text into appropriate information units or tone units. These units play an essential role for the intelligibility and auditory quality of the spoken output: they divide the flow of speech into comprehensible chunks of information, and they form the basis of prosodic rules that assign intonation contours and rhythmic structure to the spoken utterances. Without them, the strain on the listener's mind and ear will soon become intolerable.

Obviously, these 'chunking rules' - or 'segmentation rules' as I will call them - must reflect the principles of natural speech segmentation (or reading aloud) as closely as possible. A central aim of the TESS project was therefore to study these principles and try to formalize them into a set of predictive segmentation rules for automatic text-to-speech conversion (see pp 63ff). In the previous chapter (cf also Altenberg 1987a: Chapter 4), I described the correlation between tone-unit boundaries and grammatical boundaries in a popular lecture from the London-Lund Corpus and the segmentation rules that could be derived from this correlation. Here I will report on the results of a series of experiments in which these rules (somewhat adjusted) were applied to a written text from the Brown University American Corpus.

For practical reasons, the experiments were restricted to a single text from the Brown Corpus (B01: a newspaper editorial). The reason for choosing an American text - rather than, say, a British text from the Lancaster-Oslo/Bergen Corpus (see p 17) - was trivial: our speech synthesizer combined poor quality with a heavy American accent, and it did not seem quite right to expose a British text to the additional distortion of a nasal twang.

## 12.2 General principles

The segmentation is carried out by a set of rules that operate on an orthographic input text that has been automatically tagged for word classes and parsed up to phrase level (see Section 12.3 below). The segmentation rules assign tone-unit boundaries at prosodic 'breaking-points' in the grammatical structure of the text. The segmentation proceeds sentence by sentence in a series of caterpillar movements from left-to-right, during which the rules are tested in two consecutive cycles against the grammatical pattern of the sentence.

The application of the rules is determined, first, by a structural description defining the scope of each rule (the grammatical pattern to be matched by the rule), and secondly, by various constraints which impose further restrictions on the rules. The constraints are either rule-specific or general. The former, which are specified separately for each rule (see Section 12.5), define exceptions to the rules or narrow their application to certain grammatical categories or contexts. The latter (listed in Section 12.6) prevent the rules from dividing certain grammatical categories or sequences of categories. In principle, a tone-unit boundary may occur between sentences, clauses and phrases, but not inside phrases and complex words (as defined by the parser).

The location of the tone-unit boundaries reflects tendencies observed in prepared speech (see Chapter 11), although some accommodation to the conditions of a formal written text has been unavoidable. In addition, the punctuation of the input text has been used as an important cue in the segmentation process. However, since speech segmentation is to a large extent a variable and speaker-selected phenomenon, it should be emphasized that many boundaries must be regarded as optional. The treatment of these has partly been determined on the basis of informant tests, but many uncertainties remain.

The main stages in the segmentation process will be described in greater detail in Section 12.4, the individual rules in Section 12.5, and the general constraints in Section 12.6. The remaining sections will be devoted to an evaluation of the system. An example of a segmented passage is given on pp 320ff.

## 12.3 The grammatical basis: tagging and parsing

The segmentation rules presuppose an orthographic input text that has been grammatically analysed at two levels: word-class level and phrase level. These stages will not be discussed in detail here, but since the formulation and performance of the segmentation rules are dependent on the grammatical output of the tagging and parsing components, a few comments are necessary.

The tagging program makes a very detailed grammatical classification of the words in the text (see pp 101ff). This delicate classification offers obvious advantages for the formulation of the segmentation rules, which often need to refer to very specific word categories (eg subclasses of adverbs, pronouns and determiners). On the other hand, it also increases the risk of producing tagging errors, but since the tagging has generally been quite successful, there are comparatively few segmentation errors resulting from this part of the process. However, one exception should be mentioned. In an earlier version of the tagging program, regular *-ed* forms of verbs were differentiated into past (VA+D) and past participle (VA+N) forms. To facilitate the tagging this distinction was discarded in the present version (present tag: VA+DN), with the result that *ed*-forms are ambiguous. This has had unfortunate consequences in cases where a distinction between finite and nonfinite verb forms is essential (see Section 12.8.3).

The original parser developed for the London-Lund Corpus carried the grammatical analysis up to clause level, assigning functional labels (Subject, Verb, Complement, Adverbial) to phrase categories (see p 91 and Eeg-Olofsson & Svartvik 1984). A clause-level analysis was also assumed for the segmentation system explored in the previous chapter (and in Altenberg 1987a: Chapter 4). The present parser stops short at the phrase level, making a distinction between five categories: verb phrase (VPH), noun phrase (NPH), adjective phrase (JPH), adverb phrase (APH), and prepositional phrase (PPH). The segmentation rules have been adjusted accordingly: syntactic functions can no longer be identified directly, but have to be inferred from the positional patterning of the phrase categories (NPH + VPH, for example, is likely to be a Subject-Verb sequence). To some extent the delicate word classification makes up for this loss of functional information (the tag AD, for instance, uniquely identifies an adverbial disjunct), but it is inevitable that the precision and reliability of the segmentation rules have suffered somewhat from the exclusion of clause-functional categories. However, this complication of the segmentation rules must be balanced against the gain of a simplified and (presumably) more accurate parser.

Another problem is the treatment of complex phrases, ie phrases containing clauses or other phrases. Some types - especially postmodified, coordinate and appositive phrases - are not fully analysed by the parser, but merely represented as a string of structurally unrelated simple phrases:

```
from the day it convened
|__________||_||_______|
    PPH      NPH   VPH

the ban on drag racing
|_____| |____________|
  NPH        PPH
```

Consequently, there is no way in which the segmentation rules can distinguish complex phrases from sequences of simple phrases. However, there is one exception: prepositional phrases initiated by *of* and *per* (which typically function as postmodifiers) are automatically concatenated with a preceding noun phrase by the parser:

```
many other members of the Organization of American States
|___________________| |___________________| |_____________|
        NPH                   PPH                 PPH
        |_____________________|___________________|
                              NPH

$3.15 per day per patient
|___||_____| |_______|
 NPH   PPH      PPH
  |_____|_______|
       NPH
```

Since the system contains a general constraint on dividing phrases, these types are automatically treated as indivisible prosodic units by the segmentation rules. This solution works satisfactorily in the great majority of cases, but it is not infallible: *of*-constructions, though normally very cohesive, sometimes need to be prosodically divided, eg when they are discontinuous or very complex (see Section 12.8.3).

Phrasal coordinations (eg *you and I*, *the House and Senate*) also form cohesive units, but unlike *of*/*per*-constructions they are not bracketed together by the parser. Instead, special constraints on one of the segmentation rules (Rule 2) prevent these from being separated by a tone-unit boundary (provided certain structural conditions are satisfied). Thus, an example like

```
in an atmosphere of crisis and struggle
|_______________________| |__| |______|
          PPH              CA    NPH
```

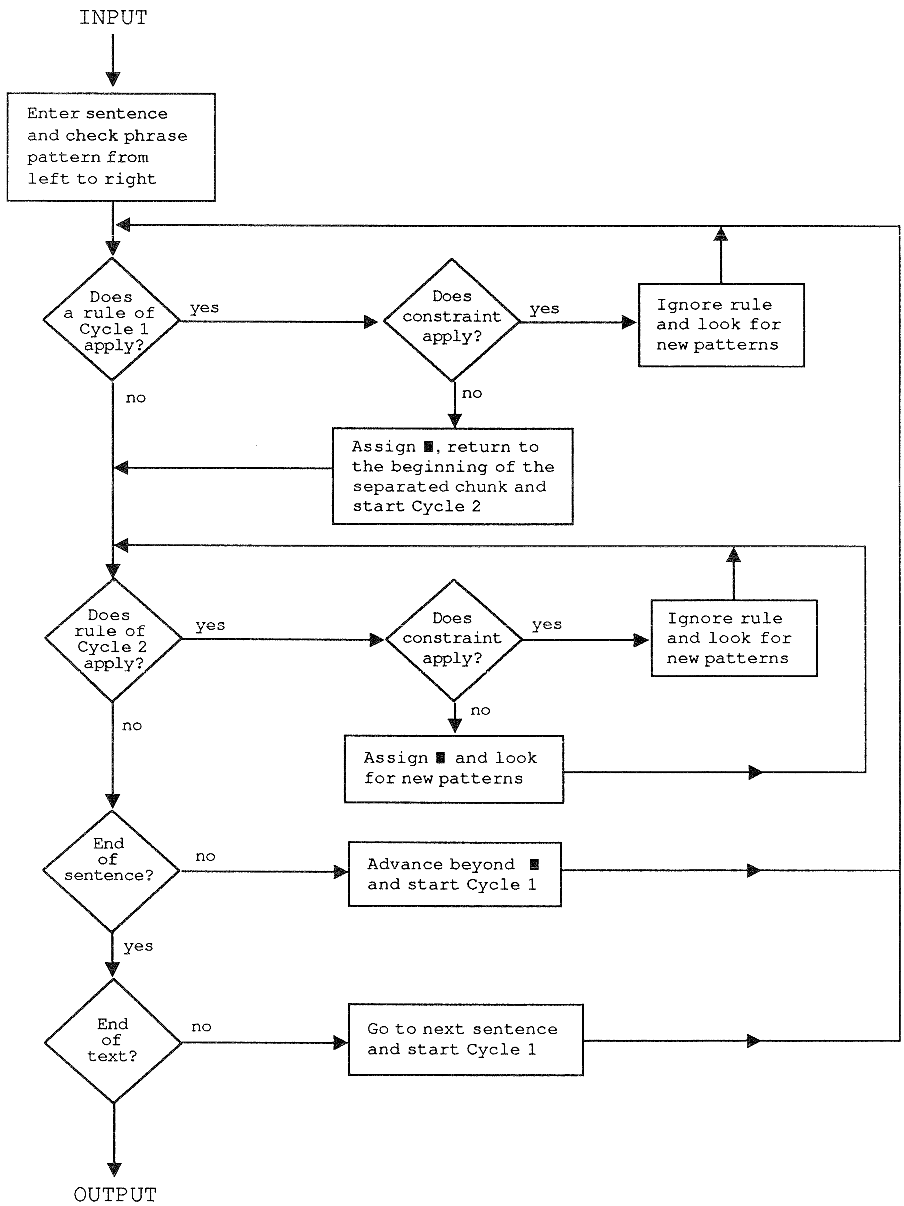
leaves the parser as a sequence of two coordinated phrases (PPH + CA + NPH), but the potential breaking-point before *and* is ignored by the segmentation rules and the entire string is produced as a prosodic unit. Not all coordinations are treated in this way, however. The exact conditions governing the segmentation of coordinations will be described under Rule 2 in Section 12.5.1.

## 12.4 Segmentation cycles

The segmentation of the text is carried out sentence by sentence by a set of eleven rules which operate in two cycles. The process, outlined in Figure 12:1, can be described briefly as follows. Each sentence is first examined for *primary breaking-points* in its grammatical structure (Cycle 1). This part of the process is handled by the first five rules of the set (Rules 1-5), which assign a tone-unit boundary at major clause boundaries and certain other junctures marked by punctuation, a coordinator, or apposition marker. As soon as a primary tone-unit boundary has been assigned in this way, the program returns to the beginning of the sentence and examines the separated segment for *secondary breaking-points* (Cycle 2). This phase is carried out by the six remaining rules (Rules 6-11), which assign a tone-unit boundary at phrase boundaries and certain lower-level (or prosodically variable) clause boundaries. When this cycle is finished, the program advances beyond the primary boundary assigned during the first cycle, re-enters Cycle 1 and starts looking for another primary breaking-point in the rest of the sentence. If such a breaking-point is located and a rule of Cycle 1 applies, a new primary tone-unit boundary is assigned; the program then returns to the beginning of the separated segment, re-enters Cycle 2 and begins to scan this segment for secondary breaking-points. This cyclic process is repeated until the whole sentence has been examined and no more rules apply. The program then advances to the next sentence and begins the same process again.

There are several reasons for this cyclic application of the rules. Statistically, prosodic breaks occur with much greater regularity at major structural boundaries than at minor constituent boundaries (see Chapter 11 and Altenberg 1987a: Chapter 4). This is a natural reflection of the fact that the former tend to separate units of greater communicative weight and independence than the latter. As a result, the prosodic separation of smaller

**Figure 12:1.** Flowchart showing outline of segmentation process.



constituents tends to be more variable and context-sensitive, being dependent on such factors as speed of delivery, emphasis, proximity of neighbouring higher-level boundaries, etc.

It is also tempting to draw a parallel between the cyclic operation of the segmentation rules and the process of speech production. When we produce utterances in speech, we use a combination of foresight and improvisation (alternating between 'top-down' and 'bottom-up' processing). Evidence from pausing and hesitation suggests that planning and execution take place at several levels, and that these levels are in some sense hierarchically organized (see Clark & Clark 1977:223ff and, for similar observations, Chafe 1979 and 1987). Even if this 'multi-tasking' process may be merged into a more or less instantaneous operation, we seem to make at least a rough 'skeleton plan' of what we intend to say before we plan and encode our ideas in detail. The amount of foresight or higher-level processing speakers have time for is obviously greater in prepared speech and reading than in spontaneous speech, and the analogy with the cyclic approach used here may therefore not be too far-fetched.

However, a cyclic application of the segmentation rules also offers decisive practical advantages. These are best illustrated by an example:

(1) The practice of charging employees for meals whether they eat at the hospital or not should be abolished. (B01 112-113)

This sentence contains at least three potential breaking-points: after *employees*, *meals* and *not*. The last two coincide with major structural boundaries and are therefore best handled in Cycle 1. The first breaking-point (after *employees*), on the other hand, occurs at a point in the clause structure (between an object and a following adverbial) where a prosodic break is common only if the final element has some communicative weight of its own and there is some distance to the next primary boundary. It is therefore more suitably deferred to Cycle 2.

The segmentation of (1) is carried out as follows. The program first enters Cycle 1 and begins to scan the sentence from left to right for a primary breaking-point. The structural description of Rule 4 (which handles breaks before non-initial adverbial clauses - see Section 12.5.1) is satisfied, and a tone-unit boundary is assigned after *meals*:

(1') The practice of charging employees for meals■ whether they eat at the hospital or not should be abolished.

The program then returns to the beginning of the sentence, enters Cycle 2 and begins to scan the separated segment for secondary breaking-points. The structural pattern underlying *charging employees for meals* matches the conditions of Rule 7 (which covers certain postverbal phrase-sequences - see Section 12.5.2), but the rule is blocked by a constraint which measures the distance to the following higher-level boundary (already assigned during Cycle 1), and no additional tone-unit boundary is assigned to the string. It should be observed here that, if the rules had not been cyclically ordered, it would have been very difficult to forestall a prosodic break at this point.

The remainder of the sentence is treated in a similar way. After two more runs of Cycle 1, additional tone-unit boundaries are assigned after *not* (by Rule 3)[1] and *abolished* (Rule 1), but no rules of Cycle 2 apply to the separated segments. The final result is:

(1") The practice of charging employees for meals■ whether they eat at the hospital or not■ should be abolished■

However, if a cyclic application of the rules is easy to justify on both theoretical and practical grounds, the distinction between primary and secondary breaking-points is less obvious. Both structurally and statistically, the difference is gradual rather than clear-cut (see Chapter 11 and Altenberg 1987a:56ff), and it would in fact be possible to argue for a segmentation process divided into three or four cycles rather than two. The division adopted here is a compromise between theoretical considerations, statistical tendencies, and practical convenience. The rules of Cycle 1 chiefly cover major clause boundaries where a prosodic break is consistently applied in speech, but also certain clause-internal boundaries where a break is less regular (eg between phrasal conjoins and appositives). Conversely, the rules of Cycle 2 mainly cover clause-internal boundaries, but also certain clause boundaries where a prosodic break is contextually determined (eg before nominal infinitive clauses and relative clauses). So far, this simple division has proved to be adequate, and there has consequently been no reason to elaborate the system.

The order of the rules within each cycle (and hence the order in which they are tested in each cyclic run) is based on their frequency of application in the text.

### 12.5.1 Cycle 1

**Rule 1.** Punctuation replacement

Pattern: . . . P . . .

Operation: replace any (word-external) punctuation mark P by a tone-unit boundary, except when P is a comma followed by

$$\left[\begin{array}{l} \textit{for example} \\ \textit{for instance} \\ \textit{however} \\ \text{NPH } \textit{said} \\ \text{etc} \end{array}\right] \quad \text{P}$$

## 12.5 The segmentation rules

In this section, the eleven segmentation rules will be presented informally, cycle by cycle. The performance of the rules will be treated in Section 12.7. The word-class symbols are explained in Chapter 3. Alternative categories are separated by a slash or listed within square brackets. In the illustrations, segmentation errors that are not relevant to the discussion have been silently corrected.

Rule 1 replaces any word-external punctuation mark[2] (indicated by 'P' in the rule) by a tone-unit boundary, except before certain metalinguistic expressions inserted as parenthetic interpolations in the text. In speech, such expressions do not normally form a tone unit of their own, but are appended with reduced pitch in the post-nuclear 'tail' of the preceding tone unit (see Crystal 1969:223f). Expressions treated in this way include vocatives, expletives and epithets (eg *John, the bastard, wouldn't lend me his car*), short comment clauses (eg *he said*, *I think*, *it seems*), and certain disjuncts (eg *actually*, *obviously*), conjuncts (eg *for instance*, *however*), and courtesy subjuncts (eg *please*, *thank you*). (For some treatments of these types of expression, see Crystal 1975:25, Allerton & Cruttenden 1976:47ff, Bing 1984, and Cruttenden 1986:43-44.) The problem with these expressions is their formal and prosodic variability. The most common and prosodically regular items can be listed, but a more adequate formulation of the rule can only be made after further research. I have here merely included a small sample of items to suggest some of the possibilities.

The effect of the rule is seen in the following examples, where all punctuation marks in the input text (a) are replaced by a tone-unit boundary in the segmented output (b), except the commas before *for instance* in (2a) and *Jefferson* in (3a):

(2) (a) This session, for instance, may have insured a financial crisis two years from now. (B01 12-14)

(b) This session, for instance■ may have insured a financial crisis■ two years from now■

(3) (a) If once they become inattentive to the public affairs, Jefferson said, you and I, and Congress and assemblies, judges and governors, shall all become wolves. (B01 40-42)

(b) If once they become inattentive to the public affairs, Jefferson said■ you and I■ and Congress and assemblies■ judges and governors■ shall all become wolves■

The ignored commas are provisionally retained in the segmented versions as a cue to subsequent intonation rules.

**Rule 2.** Before a coordinator

Pattern: $\ldots X_1$ + CA/CB/CR + $X_2 \ldots$

Operation: put ■ before CA/CB/CR unless

(a) $X_1$ and $X_2$ are words with identical tag-symbols (up to a '+' symbol, if any), or

(b) $X_1$ = BHsub, or

(c) CR + $X_2$ = *or not, or so*, or

(d) CA + $X_2$ = *and elsewhere* or *and more*, or

(e) the pattern is *between* $X_1$ + CA + $X_2$, where X may be any phrase category, or

(f) the pattern is

BHneg / BHper + CA + BHobj / *I*

If the pattern is

. . . *of* + $X_1$ + CA + $X_2$ + TO . . .

put ■ before CA irrespective of the above constraints.

Rule 2 is designed to divide clausal conjoins and separately modified phrasal conjoins linked by a coordinator, ie it will assign a tone-unit boundary to cases of the types illustrated in (4) and (5), but not to that in (6):

(4) Some other good bills were lost in the shuffle■ *and* await further action■ (B01 15-16)

(5) Only a token start was made■ in attacking the tax reappraisal question■ *and* its companion issue of attracting industry to the state■ (B01 24-25)

(6) The General Assembly■ which adjourns today■ has performed in an atmosphere of crisis *and* struggle■ from the day it convened■ (B01 1-3)

Conditions (b) - (f) prevent a prosodic separation of conjoins involving certain adverbs and pronouns. It is probable that these conditions can be simplified, but I have not explored this possibility.

It should be observed that the rule does not handle breaks between asyndetically linked conjoins or after the last conjoin in a list. The first case is generally marked by a comma in writing, and thus covered by Rule 1. For some exceptions, see Section 12.8.3, where also the second type will be discussed. Initial and medial subordinate clauses have a strong tendency to be

**Rule 3.** After initial and medial subordinate clauses.

(a) Active clause pattern:

$$\begin{array}{ccccccc} & (1) & & (2) & & (3) & \\ \ldots & \begin{bmatrix} \text{JPH} \\ \text{VPH} \\ \text{VA+G} \end{bmatrix} & + & \begin{bmatrix} \text{NP} \\ \text{PPH} \end{bmatrix} & + & \begin{bmatrix} \text{VPH-finite} \\ \text{NPH + VPH-pass} \end{bmatrix} & \ldots \end{array}$$

(b) Passive clause pattern:

$$\begin{array}{ccccccc} & (1) & & (2) & & (3) & \\ \ldots & \text{NPH} & + & \text{VPH-passive} & + & \text{VPH-finite} & \ldots \end{array}$$

Operation: put ■ after (2), if the distance from (3) to a preceding ■ is > 3 words.

prosodically separated from their superordinate clause (see Chapter 11 and Altenberg 1987a:56-58). Rule 3 is designed to handle such breaks by identifying some typical phrase patterns straddling the junction between the subordinate and superordinate clauses. Since the break may be suspended when the subordinate clause is short (less than four words), a crude distance constraint has been added to the rule to prevent the separation of such clauses. Both active and passive clause variants are provided for:

(7) *What comes after Trujillo*■ is now the puzzle■ (B01 170)

(8) [...] who is demonstrating■ *that the public trust he was given*■ was well placed■ (B01 86-90)

**Rule 4.** Before medial and final adverbial clauses.

Pattern: . . . CC . . .

Operation: put ■ before CC unless

(a) CC is preceded by CA/CB/CR, or

(b) CCcom = *as*

If the pattern is PPH + *as* + AA + *as*, put ■ before the first *as*.

Rule 4 inserts a tone-unit boundary before non-initial adverbial clauses introduced by a subordinator:

(9) Long-range planning of programs and ways to finance them■ have become musts■ *if* the state in the next few years■ is to avoid crisis-to-crisis government■ (B01 10-12)

Exceptions are made for cases in which the subordinator is immediately preceded by a coordinator (handled by Rule 2) or realized by comparative *as* (eg *such a major problem as ending the fee system*), where a break is normally suspended. In sequences like *as easily as*, a tone-unit boundary is placed before the first correlative.

**Rule 5.** Before apposition markers.

Pattern: . . . ACapp . . .

Operation: put ■ before ACapp = *for example*, *for instance*, *eg* (or *e.g.*), *ie* (or *i.e.*), *in other words*, *namely*, *such as* and *that is*.

Appositions introduced by conjuncts like *for example*, *eg*, *that is*, *ie*, *in other words*, *such as*, etc, do not always involve a higher-level break in the structure of a sentence, but have been included in Cycle 1 because of their rather strong tendency to be preceded by a tone-unit boundary:

(10) [...] he improved public facilities■ *such as* roads and sanitation■ (B01 145-148)

However, the realization of a prosodic break before (and after) some of these apposition markers is not obligatory but contextually determined. Exactly which constraints are involved is uncertain at present, and the rule must therefore be regarded as provisional.

It should also be observed that Rule 5 threatens to overlap with Rule 1, which stipulates that a comma before the mobile apposition markers *for instance* and *for example* must not be replaced by a tone-unit boundary when these are appended *after* an appositive and *followed by a punctuation mark* (see example (2) in Section 12.5.1). The distinction between initial and final position of these apposition markers (and thus between Rule 5 and Rule 1 in this respect) rests entirely on the absence or presence of a punctuation mark after the item. This is not an infallible criterion, but it will probably serve its purpose in the great majority of cases.

### 12.5.2 Cycle 2

**Rule 6.** Before nominal and relative clauses.

Pattern: . . . AB/BHitr/CD/G/TO . . .

Operation: put ■ before AB/BHitr/CD/G/TO if

(a) the distance from AB/BHitr/CD/G/TO to a preceding ■ is > 4 words or 30 characters (including spaces), and

(b) the distance from TO to a following ■ is > 3 words.

Supplementary conditions:

(c) if AB/G is preceded by PA (including *of*), put ■ before PA;

(d) if TO is preceded by *not only*, put ■ before *not*;

(e) if CD is followed by G, put ■ before G irrespective of length constraints;

(f) if G is preceded by NP or B3dem + NC, put ■ before G irrespective of length constraints.

Nominal and relative clauses are set off by a tone-unit boundary less regularly than coordinate and adverbial clauses. For various reasons (see Altenberg 1987a:59ff), their separation tends to be more dependent on the function and structure (finiteness, cohesion, degree of ellipsis) of both the subordinate and superordinate clause as well as on the proximity of higher-level boundaries. These factors have been recognized in two ways here: the rule has been placed in Cycle 2 and subjected to length constraints that make a rough estimation of the complexity of the clauses on either side of the potential boundary.

The rule applies to clauses introduced by the following items: *wh*-adverbs (AB), interrogative *wh*-pronouns (BHitr), the nominal subordinator *that* (CD), relative pronouns (G), and the infinitive marker *to* (TO). Some examples are:

(11) These women■ *whose* organization grew out of the old suffrage movement■ are dedicated to Thomas Jefferson's dictum■ *that* one must cherish the people's spirit■ but keep alive their attention■ (B01 37-40)

(12) It recognizes the fact that■ *what* helps one county■ helps its neighbors■ (B01 58-61)

The tone-unit boundary is normally placed before the clause-initiator, but may be moved before a preceding preposition (eg *in which*, *to which*) or inserted between *that* and a following relative pronoun (12). No distinction is made between restrictive and non-restrictive relative clauses (the prosodic manifestation of which is not clear-cut and here roughly handled by the length constraints), but if a relative pronoun is preceded by a proper noun or a demonstrative noun phrase, the clause is likely to be nonrestrictive and a tone-unit boundary obligatory (11).

**Rule 7.** In postverbal adverbial / complement sequences.

(a) Active clause pattern:

$$\begin{array}{ccccccc} & (1) & & (2) & & (3) & \\ \ldots & \begin{bmatrix} \text{VPH} \\ \text{VA+G} \end{bmatrix} & + & \begin{bmatrix} \text{APH} \\ \text{NPH} \\ \text{PPH} \\ \text{PA + VA+G} \end{bmatrix} & + & \begin{bmatrix} \text{PPH} \\ \text{NPH} \\ \text{PA + VA+G} \end{bmatrix} & \ldots \end{array}$$

(b) Passive clause pattern:

$$\begin{array}{ccccccc} & (1) & & (2) & & (3) & \\ \ldots & \text{NPH} & + & \text{VPH-passive} & + & \begin{bmatrix} \text{PPH} \\ \text{PA + VA+G + NPH} \end{bmatrix} & \ldots \end{array}$$

Operation: put ■ after (2), if

(a) the distance from (3) to a preceding ■ is > 4 words, and
(b) the distance from (2) to a following ■ is > 3 words, and
(c) all NPHs or PPHs in the string contain a noun head (ie N rather than BH) and APH = AA.

Rule 7 is functionally similar to Rule 6, but separates postverbal phrases rather than clauses. It places a tone-unit boundary before the second element (adverbial, direct object, complement) following a verb. As in Rule 6, the separation depends on the communicative function of the element, the complexity of the structures involved, and the proximity of neighbouring higher-level boundaries (Altenberg 1987a: 88ff). For this reason, the rule is constrained by distance conditions (backwards and forwards) in much the same way as Rule 6. Like Rule 3, it caters for both active and passive clause variants. Some typical examples are:

(13) We congratulate the entire membership■ on its record of good legislation■ (B01 29-30)

(14) Only a token start was made■ in attacking the tax reappraisal question■ (B01 24-25)

(15) It was faced immediately■ with a showdown on the schools■ an issue which was met squarely■ in conjunction with the governor■ with a decision not to risk abandoning public education■ (B01 3-5)

**Rule 8**. After clause-initial adverbials.

Pattern: ■ (C) $\left[\begin{array}{c}\text{APH}\\\text{NPH1}\\\text{PPH}\end{array}\right]$ + NPH2

Operation: put ■ before NPH2 if

(a) APH is a multi-word phrase, or a single adverb realized as

AAtim = *yesterday*, *tomorrow*, or

AC except *thus* and *such as*, or

AD except *apparently*, *certainly*, *clearly*, *maybe*, *obviously*, *of course*, *perhaps*, *probably*, *surely*, or *presumably*, or

AQpol, or

(b) NPH contains *month*, *week*, *year*, or *either way*, or

(c) PPH ≠ PA + BH.

Rule 8 separates certain clause-initial adverbials from the rest of the clause. Five types of adverbial are specified: (a) adverb phrases consisting of more than one word (eg *altogether too often*) or of certain single-word adverbs (some temporal adverbs, conjuncts, disjuncts and the politeness marker *please*), (b) adverbial noun phrases (chiefly temporal ones), and (c) prepositional phrases not containing a pronominal complement (eg *for that reason*). Some examples:

(16) *Either way*■ it sounds like a sizable hunk of money■ and is■ (B01 109-110)

(17) It was a sort of poetic justice■ that *at the time of his own demise*■ a new plot to overthrow the Venezuelan government■ [...] has been uncovered and quashed■ (B01 139-143)

(18) *Unquestionably*■ Trujillo did some good things for his country■ (B01 145-148)

The constraints on the rule are provisional and merely intended to suggest the range of phrase types that will have to be provided for. The prosodic behaviour of clause-initial phrases is still insufficiently investigated and more research is needed to clarify the main tendencies (for some useful studies, see Allerton & Cruttenden 1974 and 1978, and Stenström in Chapter 11, this volume.)

> **Rule 9**. After a complex subject.
>
> Pattern: . . . NPH + VPH-finite . . .
>
> Operation: put ■ after NPH if
>
> (a) the distance from VPH-finite to a preceding ■ is > 4 words or 30 characters (including spaces), and
>
> (b) NPH contains > 2 words, and
>
> (c) VPH is not a passive followed by PPH.

Rule 9 separates a preverbal subject from the rest of the clause, provided the subject consists of at least three words and there is a certain distance to the preceding higher-level boundary. The rule is blocked for passive clauses, which tend to be divided after the verb rather than the subject (handled by Rule 6b or 7b). Some examples:

(19) There followed the historic appropriations and budget fight■ *in which the General Assembly*■ decided to tackle executive powers■ (B01 6-7)

(20) *The intensive treatment program*■ is working well■ (B01 122-123)

> **Rule 10**. After medial prepositional phrases.
>
> Pattern: . . . PPH + VPH-finite . . .
>
> Operation: put ■ after PPH.

Rule 10 places a tone-unit boundary between a prepositional phrase and a following verb. The prepositional phrase may either be an adverbial (21) or a postmodifier of the subject (22):

(21) Long-range planning of programs [...]■ have become musts■ if the state *in the next few years*■ is to avoid crisis to crisis government■ (B01 10-12)

(22) Confidence *in the state's economic future*■ is reflected in the Georgia Power Company's record construction budget■ for this year■ (B01 97-99)

In the latter case, the rule acts as a useful supplement to Rule 9 (for occasional overlap with Rule 3, see Section 12.9).

> **Rule 11.** In sequences of three prepositional phrases.
>
> Pattern: . . . PPH1 + PPH2 + PPH3 . . .
>
> Operation: put ■ after PPH2.

When more than two prepositional phrases occur in sequence, (at least) the third tends to be prosodically separated from the first two:

(23) Even *with the increase in funds*■*for the next fiscal year*■ Georgia will be spending only around $3.15 per day per patient■ (B01 116-118)

(24) Kansas■ regarded *as tops in the nation*■ *in its treatment of the mentally ill*■ spends $9 day per patient■ (B01 120-121)

## 12.6 General constraints

Apart from the rule-specific constraints restricting the application of individual rules, there is need for certain general constraints to prevent the separation of word sequences that tend to function as cohesive units. These sequences are of three main kinds:

(a) combinations analysed by the parser as syntactic (simple or complex) phrases (with the single exception specified in Rule 2), eg *exactly how far* (APH), *very real* (JPH), *the usual spate of silly resolutions* (NPH), *in the next few years* (PPH), *has been taken* (VPH);
(b) contractions (*it's, I'd, they'll, doesn't*, etc);
(c) various other sequences, most of which involve a 'light' first element such as a pronoun or closed-class adverb (see Altenberg 1987a:86ff).

Only the last category needs to be specified here. Six subtypes can be distinguished:

(c1) Pronoun-verb sequences: AB/BH (except BHobj) + V

Examples: *it (nobody, anything, mine, how) is*
*you (some, these, many, others) are*

(c2) After certain 'light' elements:

| | | |
|---|---|---|
| AS | BHsub*V | + X |
| AX | | |
| BHdem*V | G | |
| BHitr | GAwho*V | |
| BHitr*V | GCwha*V | |
| BHneu*V | GDtha*V | |
| BHper*V | *as* | |
| BHsub | | |

Examples: *much (slightly, greatly, really) amused*
*there (she, I, who, what) certainly was*
*it's (that's, you've, she'd) probably*
*as quickly (as)*

This constraint does not apply to boundaries coverd by Rule 1.

(c3) Before certain 'light' adverbs: AN, AE

Examples: *could not, good enough*

(c4) After nonfinite verb phrases: VPHnonfin

Examples: *accepting her, arriving in London*

(c5) Certain indefinite pronouns + prepositional *but*:

BHneg / BHnon / BHuni + *but* (= PA)

Examples: *nothing (anything, all) but*

Most of these constraints are tentative. For example, the restriction in (c3) on placing a tone-unit boundary before AN (*not*) may have to be restricted to cases in postverbal position. Moreover, some of the constraints may be (wholly or partly) redundant. For example, it is doubtful if any rule in the present system threatens to separate a pronoun or other light element from a following element, or a preposition from a following nonfinite verb. This potential redundancy is a remnant from an earlier version of the system, which needed rather extensive safeguards against a 'blind' default rule which served to divide overlong strings produced by the system. In a more definitive system this redundancy must be eliminated, but at an experimental stage it may be useful to retain some safeguards against undesired results.

It should be added that, apart from these constraints, there is no general restriction on tone-unit length in the present system. In most speech varieties tone units tend to be fairly short: their mean length is normally 4-5 words and few tone units exceed seven words (see Altenberg 1987a:23ff).[3] It would therefore seem justified to restrict the tone units produced by the system to a maximal length of, say, 7-8 words. However, one problem with this solution is that, in speech converted from written text, tone units sometimes have to be longer (in the present text the average length of the tone units produced by the system is 5.5 words). The tolerance limit for unbroken tone units seems to be around 10 words, but tone units containing complex noun phrases occasionally have to be longer:

(25) [...] the 13 northwest Georgia counties■ *that are members of the Coosa Valley Area Planning and Development Commission*■ (B01 49-52)

(26) Raphael Trujillo■ *the often blood-thirsty dictator of the Dominican Republic for 31 years*■ perhaps deserved his fate [...] (B01 129-131)

In long tone units of this kind, where the program fails to detect a natural breaking-point, any further division imposed by default is bound to be both arbitrary and unnatural. The problem has therefore been left unsolved here.[4]

## 12.7 The performance of the rule system

The performance of the segmentation system can be evaluated in at least two ways. We can either look at the boundaries that the rule system assigns to the text and determine their appropriateness, or we can determine what prosodic boundaries are needed in the text and see how well the system matches this ideal segmentation. Both approaches presuppose a pre-established norm against which the system is compared, but the second has the advantage of revealing not only how well the rules perform when they apply, but also what boundaries they fail to assign. I will therefore use the second approach here. The evaluation is based on an intuitive reading of the text (which in turn reflects tendencies observed in prepared speech), but doubtful cases have been specially tested with a group of informants who were asked to read the text as a radio/TV news bulletin.

The performance of the rules is shown in Table 12:1 (for an illustration, see also pp 320ff).

**Table 12:1.** Performance of the segmentation rules. Errors caused by external factors (failures in tagging, parsing and neighbouring rules) are excluded.

| RULE | + | ? | - | TOTAL |
|---|---|---|---|---|
| CYCLE 1 | 221 (94%) | 6 (3%) | 7 (3%) | 234 |
| 1 | 179 | 0 | 1 | 180 |
| 2 | 23 | 6 | 2 | 31 |
| 3 | 9 | 0 | 3 | 12 |
| 4 | 7 | 0 | 0 | 7 |
| 5 | 3 | 0 | 0 | 3 |
| X | - | 0 | 1 | 1 |
| CYCLE 2 | 69 (64%) | 20 (19%) | 18 (17%) | 107 |
| 6 | 24 | 4 | 1 | 29 |
| 7 | 19 | 5 | 6 | 30 |
| 8 | 9 | 1 | 1 | 11 |
| 9 | 7 | 2 | 0 | 9 |
| 10 | 7 | 0 | 0 | 7 |
| 11 | 3 | 1 | 0 | 4 |
| X | - | 7 | 10 | 17 |
| TOTAL | 290 (85%) | 26 (8%) | 25 (7%) | 341 |

The rules are listed in order of application, with those of Cycle 1 preceding those of Cycle 2. Boundaries not covered (and hence missed) by the system are grouped together and marked 'X' in the table. For each rule is indicated the number of appropriate (+), doubtful (?) and inappropriate (-) boundary assignments made by the rule. Doubtful cases involve boundaries judged to be optional or uncertain.

Two types of rule failure have been excluded in the table: those due to a tagging or parsing error, and those due to failure of a preceding rule.[5] Although such rule-external errors provide useful information about the overall efficiency of the system, they tell us nothing about the performance of individual rules and will therefore be disregarded here.

As shown in Table 12:1, the overall failure rate of the system is fairly low (7%), and the number of doubtful applications comparatively small (8%). If we regard the latter as acceptable - and in slow speech they generally are - the success rate of the system as a whole is well above 90%, a result that must be regarded as very satisfactory.

Most of the rules have a very low error tendency: five produce no errors at all (Rules 4-5 and 9-10) and four only occasional errors (Rules 1-2 and 6 and 9). Only two rules have a failure rate of 20-25% (Rules 3 and 7).

If we compare the two cycles, we find that the rules of Cycle 1 are on the whole more successful (3% failure) than those of Cycle 2 (17% failure). This partly reflects the fact that most of the 'X-cases' (missing boundaries not covered by the rule system) have been interpreted as belonging to Cycle 2, but it is significant that the rules of Cycle 2 also produce a greater number of doubtful cases than those of Cycle 1. In other words, the boundaries covered by Cycle 2 tend to be less regular and predictable than those of Cycle 1, and the rules designed to handle them tend to be more difficult to define or formalize. This obviously reflects the greater indeterminacy and context-dependency of the lower-level boundaries covered by Cycle 2. However, to understand the reasons for rule failure and determine how the system might be improved, it is necessary to examine the rules in greater detail.

## 12.8 Reasons for rule failure

To simplify matters I will concentrate almost entirely on rules producing unacceptable boundaries, while doubtful cases will largely be ignored. Errors not covered by the rule system will be discussed in Section 12.8.3. The following abbreviations will be used:

| | | |
|---|---|---|
| * | = | inappropriate tone-unit boundary |
| [+] | = | missing but desired tone-unit boundary |
| [?] | = | missing but doubtful tone-unit boundary |

### 12.8.1 Cycle 1

*Rule 1. Punctuation replacement*
Rule 1 is the most frequently applied and reliable rule in the system (1% failure). It successfully handles nearly half of the boundaries in the text, failing only in one case involving a vocative:

(27) Chin up* Soapy■ (B01 96)

The problem with vocatives is that they are often formally indistinguishable from non-restrictive appositions: both typically consist of a proper noun preceded by a comma. However, while non-restrictive appositions normally follow a noun phrase (eg *the Secretary of State, Dean Rusk*), vocatives less often do. This difference might be used to improve the existing rule, for example by adding the pattern [... X, NP] (where X must not be a noun phrase) as a further exception to the rule.

*Rule 2. Before a coordinator*

Although coordination is a very complex phenomenon, involving a range of formal variants (see Quirk et al 1985: Chapter 13), Rule 2 works surprisingly well (6% failure). Only phrasal coordinations occasionally cause problems:

(28) [...] you and I■ and *Congress* and assemblies*■ judges and governors■ shall all become wolves■ (B01 40-42)

(29) [...] Secretary of State Rusk junks *bluff* and nuclear brinkmanship*■ and builds more muscle■ and greater safety■ into our military position■ (B01 75-78)

In (28), *Congress and assemblies* must be produced as a single phrase to preserve the parallelism of the list. The conjoins are unmodified, but the rule fails here because the members realize different word classes, proper and common noun. The rule can easily be changed to accommodate noun pairs of this kind, but such a change is not without risks (cf *girls love John*■ *and children adore him*).

In (29), the second conjoin (*nuclear brinkmanship*) is separately modified, but the first (*bluff*) is hardly 'heavy' or semantically distinct enough to justify a separation. Exceptions of this kind are difficult to capture in purely formal terms, however, and I have therefore left them unsolved here.

In addition to these comparatively clear cases of rule failure, there are some intermediate examples in the text where a tone-unit boundary is optional or doubtful (eg after *muscle* in (29)). However, I will disregard them here.

*Rule 3. After initial and medial subordinate clauses*

Although Rule 3 operates satisfactorily in the majority of the cases, it has a comparatively high failure rate (25%). The errors are all due to 'irregularities' in the structural pattern, which mislead the rule to identify a non-existent clause boundary:

(30) Dekalb budget shows county* is on beam■ (B01 79-80)

(31) The harder the choice■ the more willing the league* is to wade in■ (B01 46-47)

In (30) - a headline - the true clause boundary is concealed by ellipsis of the nominal clause subordinator *that*; in (31) marked word order is the misleading factor.

Deviant clause patterns of this kind are a constant threat to the generality and descriptive economy of the rules. They also raise interesting questions about the division of labour between different components of the system and the cost of elaborating one part at the expense of another. Since it is difficult to see how Rule 3 might be improved in a simple way, the best solution seems to be to refine the parser to eliminate structural ambiguities of the kinds illustrated above, at least that represented in (30). I will return to this possibility in Section 12.8.3, where a similar type of rule failure (caused by ellipted relative pronouns) will be discussed.

### 12.8.2 Cycle 2

*Rule 6. Before nominal and relative clauses*
Despite the contextual sensitivity of this rule, it normally produces satisfactory results (3% failure). Some typical problems are illustrated in the following examples:

(32) The recent history of the Dominican Republic■ is an almost classical study[+] of *the way* in which* even a professedly benevolent dictatorship■ tends to become oppressive■ (B01 143-145)

(33) There was considerable evidence■ of a tacit rapprochement with Castro in Cuba■ previously a bete noire to Trujillo■ thus illustrating *the way*[?] *in which* totalitarianism of the right and left coalesces■ (B01 167-170)

Both examples contain the sequence *the way in which.* The rule prescribes a break before *in which* unless there is already a tone-unit boundary in the immediately preceding context. Consequently a tone-unit boundary is assigned to the example in (32) but not to that in (33), although our intuitive preference may be the reverse. Example (32) can be remedied if we make special provisions for the sequence *of the way in which*, requiring the break before *of* rather than *in* (thus violating the constraint on dividing *of*-constructions). Example (33) might be saved by lifting the distance constraint, but this will have serious consequences for a number of other examples. Hence both examples are rather special cases, and neither can be salvaged without violating general and otherwise useful constraints.

Example (33) also illustrates another problem: the difficulty of chunking a written text in a natural way. The test panel divided (33) as follows (the figures in parentheses indicate the percentage of test subjects preferring a tone-unit boundary at that point):

(33') [...] thus illustrating (23%) the way (31%) in which (8%) totalitarianism (15%) of the right (8%) and left (23%) coalesces.

The responses show no clear majority for a prosodic division anywhere in the string. Moreover, since less than a third of the test panel preferred a break after *way*, it is difficult to justify a tone-unit boundary there. Yet, without a prosodic break somewhere in the string the resulting tone unit will be longer than is normally tolerated in speech (see Section 12.6). Since a break after *way* is after all judged to be the best alternative, we may regard it as acceptable in the circumstances.

*Rule 7. Postverbal adverbial/complement sequences*

Any written text is likely to abound in postverbal phrases, and Rule 7 is consequently a frequently tested rule in the system. Considering this, its comparatively high failure rate (20%) is not very surprising. Most of the errors are due to the crudely formulated distance constraints which sometimes block the rule unnecessarily. Two examples are:

(34) The League of Women Voters [...]■ is inviting financial contributions[+] in the wind-up of its fund drive■ (B01 34-36)

(35) This left the Soviets plenty of leeway■ to start low-grade brushfire aggressions[+] with considerable impunity■ (B01 73-75)

To measure communicative 'weight' or complexity in terms of graphic words is of course a mere practical convenience. In a full-fledged text-to-speech system, where phonological units must be computed, the use of phonemes or syllables would no doubt serve the purpose better, but such an approach has been out of the question here. However, earlier versions of the rule system have shown characters to be a more delicate criterion than words, and a supplementary distance condition of this kind has been added to Rule 6 with good effect. This can be seen in the following example, where the matrix clause is only four words long, yet long enough in terms of characters to justify a separation of the following clause:

(36) Thereupon followed a demonstration■ that tyranny knows no ideological confines■ (B01 163-164)

The absence of a similar condition in Rule 7 is an unfortunate oversight. A limitation of the distance constraints (backwards and forwards) to, say, ≥ 25 characters (including spaces) would in fact salvage all the failures of Rule 7 (with the exception of one case, to be discussed below). Whether this method can replace the word criterion altogether, and whether it will have any undesired side-effects on the application of the rule, has yet to be determined.

In other respects, Rule 7 operates quite satisfactorily. It captures what appears to be a very regular tendency in the segmentation of postverbal structures: when a division is required, it normally occurs after the first element following the verb, no matter what the phrase pattern is. (Indeed, this tendency is so consistent that the specificaiton of the rule might be simplified to: 'put ■ after the first phrase following VPH'.) This obviously reflects a basic feature in the structure of clauses: the element immediately following the verb (normally an object) tends to have a more central function than any subsequent elements (normally adverbials). There is only one exception to this tendency in the text:

(37) Coupling its own budget of $83,750■ with a $30,000 state grant [...]■ the group expects to sign *a contract* in March*[+] *with Georgia Tech.*■ (B01 52-55)

In this example, where an object is followed by two adverbial phrases, the natural breaking point is after (rather than before) the first of these, ie the obvious grouping is to link the time adverbial *in March* with the predicate rather than the following adverbial. This grouping, which is semantically justified, can hardly be predicted on formal grounds, and I shall leave it without any further suggestions.

*Rule 8. After a clause-initial adverbial*

Rule 8 is normally successful (9% failure), but fails in the following instance involving an initial *wh*-adverbial:

(38) But exactly *how far** it will go toward improving conditions■ is another question■ (B01 110-112)

Since clause-initial position is obligatory with interrogative *wh*-elements, there is normally no communicative need to separate it from the rest of the clause. This special thematic status of *wh*-interrogatives may be recognized by blocking Rule 8 in cases where an initial APH contains a *wh*-adverb (AB).

### 12.8.3 Some unsolved problems

The errors discussed so far have all illustrated failures in the operation of the existing rules of the system. But there is also another group of errors which falls outside the scope of the present rules (indicated by X in Table 12:1). This group consists of cases where a prosodic break is needed but no rule exists to take care of it. Although these errors are not very numerous, they are important because they indicate some of the problems that remain to be solved. I will briefly discuss the main cases here.

**Breaks after phrasal conjoins**

Rule 2 separates various types of syndetic coordination, but does not assign a prosodic break *after* the last conjoin in a sequence. Such a break is often required, especially if the final conjoin is separately modified (cf Crystal 1975:19):

(39) That such expansion can be obtained■ without a raise in taxes■ is due to growth of the tax digest■ *and sound fiscal planning*[+] on the part of the board of commissioners■ (B01 86-90)

(40) More attendants■ nurses *and doctors*[+] should be hired■ (B01 115)

A boundary in this position can be achieved fairly easily by a special rule or by an addition to Rule 2. Whether a break is also needed between the last two elements in a list of conjoins (*A, B and C*), even in unmodified cases of the type illustrated in (40) (ie *nurses*[?] *and doctors*), is more uncertain.[6] The formulation of such a rule is also more problematic, since it would violate the principles of Rule 2 (which does not permit a break before unmodified conjoins). I have no ready solution to this problem.

**Breaks in *of*-phrases**

A more serious problem is illustrated in the following examples:

(41) Certainly all can applaud passage of an auto title law■ [...] acceptance by the state[+] of responsibility for maintenance of state roads■ in municipalities at the same rate as outside city limits■ (B01 16-21)

(42) There was considerable evidence[+] of a tacit rapprochement* with Castro in Cuba■ previously a bête noire to Trujillo■ (B01 167-170)

(43) The league of workers search out the pros and cons[+] of the most complex issues■ and make them available to the public■ (B01 44-46)

In all these examples a prosodic break is needed before a postmodifying *of*-phrase. Breaks in complex phrases are not of course unusual as such (see Altenberg 1987a:107ff), but they tend to be less common before *of*-phrases, which is one reason why these have been analysed as part of complex phrases by the parser in the present system, and subsequently prevented from separation by a general constraint on phrase splitting. Exceptions to this constraint are of course possible, but the problem is when such exceptions should be permitted. There are no formal indications in the examples above to determine this (except the very length of the undivided strings).

**Breaks in other postverbal strings**

The problem illustrated in (41-43) is not restricted to *of*-constructions but threatens any string of prepositional phrases in postverbal position. Such strings are normally handled by Rule 7 or 10, but these cannot cover all possible variants:

(44) [...] an issue which was met squarely■ *in conjunction with the governor*[+] *with a decision* not to risk abandoning public education■
(B01 3-5)

(45) Georgia's mental health program■ received a badly needed boost■ *from the General Assembly*[+] *in the form of a $1,750,000 budget increase*■ *for the Milledgeville State Hospital*■ (B01 104-106)

(46) The NCTA is well advised■ to seek funds *for this purpose*[+] *from the present session of Congress*■ (B01 192-193)

In (44), *in conjunction with* is tagged as a complex preposition, and the phrase pattern consequently fails to satisfy Rule 11. A possible solution in cases of this kind is to allow Rule 11 to take account of complex prepositions with an underlying PA + NC + PA structure, but this possibility must be tested further.

In example (45), the verb *received* is followed by a long string of phrases (NPH + PPH + PPH + PPH) divided at two points: after *boost* (by Rule 7) and *increase* (by Rule 11). However, an equally important break is needed after *Assembly*. A possible way of handling strings of this kind is to supplement Rule 11 with a length condition which allows an extra break between the first two prepositional phrases (in a sequence of three) in cases where an unbroken sequence would be too cumbersome.

Example (46) also contains a long postverbal string of phrases (after *seek*), with the pattern NPH + PPH + PPH. Rule 7 is constrained (by a leftward distance condition) and the desired break therefore has to be postponed one step to set off the second (rather than the first) prepositional phrase. There is no rule to cover this case, but a possible solution would be to add a late backup rule to take care of postverbal strings left undivided by Rule 7 in cases where the latter is blocked by a leftward constraint. Alternatively, such a rule could be generalized to take care of all three examples by inserting a break after the first phrase in an unbroken string of prepositional phrases.

**After postmodifying adverbs**

Postmodifying adverbs are not analysed as postmodifiers by the parser but represented as independent adverbial phrases. Hence, in the following examples there is no formal indication that the underlined adverbs are part of a complex subject, and Rule 9 consequently fails in both cases:

(47) Newspapermen and politicians *especially*[+] are aware of the penetrating attention■ and expert analysis■ the league gives to public affairs■ (B01 42-44)

(48) Because the buses would not stop on the parkway■ land for bus stations■ and for parking areas *nearby*[+] will be needed■ (B01 190-192)

Though the difficulty caused by these adverbs is strictly a parsing problem, it is of interest here because it is connected with the segmentation of medial adverbials generally. Both adverbs above occur before a finite verb phrase, and they are therefore potentially within the scope of Rule 10, which separates a medial adverbial from a following verb (on the overlap between Rules 9 and 11, see Section 12.9). At present, Rule 10 is restricted to adverbials realized by prepositional phrases, but it is obvious that many adverbs also require a following tone-unit boundary, especially when they function as emphatic adjuncts and subjuncts (see Altenberg 1987a:97ff). If Rule 10 is extended to include such adverbs, it is possible to cover postmodifying adverbs of the type illustrated in (47) and (48). However, our knowledge of the prosodic tendencies of medial adverbs is still too limited for any definite suggestions.

**Before ellipted clause initiators**

In each of the following examples a prosodic break is required before a subordinate clause (nominal and relative respectively):

(49) It also weakened our diplomatic stance■ because Russia could easily guess[+] we did not desire a nuclear war■ except in the ultimate extremity■ (B01 71-73)

(50) [...] considering the additional half-million dollars[+] Gov. Vandiver allocated last year■ from the state surplus■ (B01 106-109)

Breaks before nominal and relative clauses are normally handled by Rule 6, but this rule applies only if a clause initiator (subordinator or pronoun) is overtly realized. Cases where the clause-initiator is ellipted, as in (49) and (50), are rare in the text and have not been catered for in the rules, but in a more refined system they will have to be handled in some way, preferably by an improved parser, but alternatively by elaborating the pattern of Rule 6.

In the following example a break is also needed before a nominal clause, but what is 'ellipted' here is not a subordinator but a quotation mark (and/or comma) signalling the beginning of direct speech:

(51) Then he arrived in Zanzibar■ and found Africans carrying signs saying[+] American imperialists■ go home■ (B01 94-96)

Without the aid of some orthographic clause-marker (which would make Rule 1 apply correctly), cases of this type are difficult to handle.

**Other postverbal phrase patterns**

There remain two deviant patterns to account for, both broadly within the domain of Rule 7:

(52) Cities and counties interested in industrial development■ would do well in the months ahead■ to keep their eyes* *peeled*[+] toward the 13 northwest Georgia counties■ that are members of the Coosa Valley Area Planning and Development Commission■ (B01 49-52)

(53) If once they become *inattentive*[?] to the public affairs, Jefferson said■ you and I■ and Congress and assemblies■ judges and governors■ shall all become wolves■ (B01 40-42)

Example (52) contains an object-with-participle construction with a prosodic break erroneously placed before (rather than after) the participle *peeled*. The reason for the error here is the ambiguous tag (VA+DN) assigned to *peeled*, which allows Rule 3 to apply despite its constraint on a nonfinite verb phrase after the tone-unit boundary. But even if the verb had been tagged unambiguously as nonfinite, a rule is needed to place the boundary after *peeled*. This could be achieved by adding the pattern [VPH + NPH + VA+N■ PPH] to Rule 7.

Example (53) seems to require a break after *inattentive*. This can easily be achieved by adding JPH to the pattern of Rule 7, but since the break seems to be optional, this possibility must be tested further.

## 12.9 Redundancy in the rule system: overlapping rules

Since the segmentation rules have to match a great variety of structural patterns (see eg Rules 3 and 7), it is inevitable that some of the rules should overlap. Some redundancy in the system is probably unavoidable, and may not be altogether harmful, but as it is likely to affect the efficiency of the system in one way or another, I will briefly examine its effects here.

Since the rules of the system operate in a fixed sequence (see Section 12.5), partial overlap between the rules may have two consequences:

(a) an earlier rule may forestall a later one;
(b) a later rule may replace an earlier (failing) rule.

Neither of these possibilities turns out to be very common in the present text, however, and fortunately perhaps, (a) is less common than (b). In fact, there is only one clear instance of (a):

(54) Then a full-time planning office■ will be established in Rome■ to work with a five-member Georgia Tech research staff■ (B01 55-58)

Here the assignment of a tone-unit boundary after the complex subject *a full-time planning office* is properly the task of Rule 9 (Cycle 2), but Rule 3 (Cycle 1), which serves to separate initial subclauses, applies instead. This substitution is not entirely accidental (the rules have similar functions and partly similar patterns), but the intended result is undoubtedly achieved by the wrong rule. The example is a rare illustration of how a parsing error may be 'repaired' by rule substitution.

Case (b), which allows a later rule to act as a backup for an earlier one, is more common (7 instances). It is restricted to a small set of partially overlapping rules: Rule 8 may substitute for Rules 3 or 6, and Rule 10 for Rules 3 or 9. The substitution is normally possible only when the earlier rule has failed to match a complex clause-initial structure, ie it is reminiscent of the situation in (54), although the substitution is reversed and likely to have a safer outcome. Two examples will suffice to illustrate this. In (55) and (56) below, the assignment of a tone-unit boundary after *effort* and *future* is the task of Rules 3 and 9 respectively. However, for various reasons these rules fail and the boundaries are assigned by Rules 8 and 10 instead:

(55) [...] and that by banding together■ in an area-wide effort■ better results can be accomplished■ than through the go-it-alone approach■ (B01 58-61)

(56) Confidence in the state's economic future■ is reflected in the Georgia Power Company's record construction budget■ for this year■ (B01 97-99)

These examples illustrate clearly the positive value of partial rule overlapping: if one rule fails, another is there to take over its function. This is the prevailing effect of this type of redundancy in the system. I have only noticed one situation where its effect may be harmful. This situation arises in cases where an earlier rule is blocked by a distance constraint but subsequently overridden by a rule without such a constraint. This is illustrated in (57), where Rule 8 has separated the restrictive relative clause after *prisoners* in spite of the fact that Rule 6 (which handles relative clauses) is blocked:

(57) The jails were filled to overflowing■ with political prisoners■ who had incurred his displeasure■ (B01 150-152)

However, with this single exception (where the segmentation is uncertain anyway), the slight redundancy that exists in the rule system appears to have a predominantly beneficial effect on the segmentation. Hence, there has been no reason to sharpen the rules in this respect.

## 12.10 Conclusion

Natural speech segmentation is determined by a variety of factors - cognitive, semantic, pragmatic, emotional, grammatical and communicative. Considering the elusiveness of these factors, it is surprising that automatic simulation of speech segmentation can be achieved at all. That this is not only possible, but can in fact be done fairly well, can be ascribed to the mediating role that grammar has in the segmentation process, a role that is emphasized in text-to-speech conversion, where the starting-point is a written text. By basing a segmentation program on the grammatical structure (and punctuation) of a text, it is possible to make a crude 'interpretation' of the text and chunk it prosodically in a rather satisfactory way.

The segmentation program described here has a success rate of over 90% (discounting doubtful cases and 'system-external' errors). Although this must be regarded as a very successful result, it is obvious that many features of the system are still preliminary and that many improvements can be made. I will therefore conclude with a few general observations on the program and point out some areas where refinements are needed.

The most important factors affecting the performance of the segmentation program are:

(a) the parser (word-class tagging and phrase analysis);
(b) the segmentation rules;
(c) the correspondence between prosodic segmentation and grammatical structure;
(d) the influence of situational factors (type of input text, speed of delivery, etc).

The grammatical analysis provided by the parser is obviously of crucial importance to the segmentation. An evaluation of the parser is not my concern here, but since the segmentation program offers valuable feedback on the validity of the grammatical analysis, a few comments may be justified.

The parser used here analyses the input text in two consecutive steps: word-class tagging and phrase category assignment. On the whole, these two levels provide a sufficient basis for the segmentation rules. The absence of a third, clause-functional level (identifying categories like Subject, Object, Adverbial, etc) complicates the segmentation in some respects, but the consequences are less serious than might be expected. As the results demonstrate, clause-functional categories are generally dispensable provided that the analysis at the lower levels is accurate and sufficiently delicate.

The detailed, semantically-oriented word-class tagging does much to compensate for the absence of a clause-functional level, especially in the case of adverbials. The tagging is generally adequate, the main exception being its failure to disambiguate verbal *ed*-forms, which play a crucial role for the identification (and potential separation) of finite and nonfinite clauses (Rules 3, 9, 10).

The phrase analysis is also generally adequate, but improvements are needed in the treatment of complex phrases (especially postmodified and coordinated phrases), the interpretation of *ing*-words (as verbs, premodifiers or noun-phrase heads), and the identification of clause boundaries concealed by ellipted subordination markers. Some of the problems created by these constructions can be solved by elaborating the segmentation rules (as has partly been done in the case of coordinations), but much would be gained if this could be done by the parser. However, the optimal division of labour between the components can only be determined after further testing.

The segmentation system as such depends on three main features: the structural definition of the rules (including the rule-specific constraints), the general constraints, and the cyclic application of the rules. Of these, only the first creates any problems worth considering here. When a segmentation error is produced, the reason is either that one of the existing rules has failed or, less commonly, that no rule has been formulated to handle the case. Some of these errors can be eliminated fairly easily by minor adjustments of the existing rules or by an extension of the system to cover missed boundaries. Many cases, however, require further research and testing before any improvements can be made. In particular, more research is needed to determine the separability of initial and medial adverbials (Rules 1, 8 and 10), of phrasal conjoins (Rule 2), and of certain context-sensitive clause types (Rules 3, 6, 7 and 9).

The context-dependency of many boundaries emphasizes the variable nature of prosodic segmentation. Some of this flexibility has been possible to capture by a combination of cyclic rule-ordering and context-sensitive constraints which make a rough estimation of the complexity of constituents and the distance to higher-level boundaries. However, these measures are not sufficient to handle the influence of situational variables such as communicative purpose and speed of delivery. More research is therefore needed to test the acceptability of automatically segmented texts under different situational conditions. Speech rate, in particular, seems to interact with some of the context-sensitive rules (especially those of Cycle 2), and an important task for future research will be to determine the extent of this interaction and the possibility of incorporating some kind of 'speed index' into the system.

## Notes

1 In the present version of the system, Rule 3 does not in fact assign a tone-unit boundary after *not*, since the segmentation of coordinations is only partly solved by Rule 2 (see the discussion in Sections 12.5.1 and 12.8.3). However, this inadequacy can be eliminated by an addition to Rule 3 or, better, Rule 2 (where *or not* is already specified).

2 By 'word-external' punctuation mark is here meant all punctuation marks except apostrophes and hyphens. The treatment of quotation marks and parentheses (and other types of bracketing) is uncertain, but since none of these occurred in the experimental text, they have been ignored here.

3 In a special study of a TV news program from the files of the Survey of English Usage (text W.2.4), the mean length of tone units (including subordinate tone units) was found to be 4.4 words in news reporting and 4.9 words in news reading. In other words, the average tone-unit length does not seem to vary much from one functional variety to another.

4 One way of reducing the average tone-unit length produced by the system is to relax the distance constraints of some of the rules. Complex subjects (Rule 9), in particular, can easily be made more 'detachable' in this way. However, the separability of various constituents is closely connected with speech rate, and it is difficult to make any definite suggestions until more is known about this relationship.

5 The failure of an earlier rule in the system sometimes affects the distance constraints of a later rule, causing the latter to fail as well. It is this latter type of rule failure that is disregarded in Table 12:1. If all rule-external errors are counted, the failure rate of the system rises to 13%. Most of the tagging and parsing errors involve *-ed* and *ing*-forms of verbs.

6 Example (40) was not included in the judgement test, but other cases of the same type were tested. The results suggest that a tone-unit boundary before the last conjoin in a list may be optional:

(a) ... studying (15%) Georgia's (8%) very real (38%) economic, (46%) fiscal (38%) *and social problems* (100%) and come up with answers ... (B01 30-34)

(b) ... essential services (85%) such as health and welfare, (100%) fire protection, (100%) sanitation (69%) *and road maintenance.* (100%) (B01 82-86)

## A segmented sample passage

To give an indication of the performance of the segmentation program, an extract from the experimental text (an editorial from *The Washington Post*, June 1, 1961) is given here for illustration. The original input text is presented first, followed by the segmented version. To facilitate a comparison, the sentences have been numbered in both versions. (Note that, due to a flaw in the program, sentence (5) was never segmented and is therefore missing in the segmented version.)

Apart from some orthographic normalization, no changes or correction of segmentation errors (including those due to tagging or parsing failure) have been made in the segmented version: the output is presented exactly as is leaves the program.

Input text:

(1) *The end of Trujillo*

(2) Assassination, even of a tyrant, is repulsive to men of good conscience. (3) Rafael Trujillo, the often blood-thirsty dictator of the Dominican Republic for 31 years, perhaps deserved his fate in an even-handed appraisal of history. (4) But whether the murder of El Benefactor in Ciudad Trujillo means freedom for the people of the Caribbean fiefdom is a question that cannot now be answered.

(5) Trujillo knew a great deal about assassination. (6) The responsibility for scores of deaths, including the abduction and murder of Jesus Maria Galindez, a professor at Columbia University in New York, has been laid at his door. (7) He had been involved in countless schemes to do away with democratic leaders in neighboring countries such as President Romulo Betancourt of Venezuela. (8) It was a sort of poetic justice that at the time of his own demise a new plot to overthrow the Venezuelan government, reportedly involving the use of Dominican arms by former Venezuelan Dictator Marcos Perez Jimenez, has been uncovered and quashed.

(9) The recent history of the Dominican Republic is an almost classical study of the way in which even a professedly benevolent dictatorship tends to become oppressive. (10) Unquestionably Trujillo did some good things for his country: he improved public facilities such as roads and sanitation, attracted industry and investment and raised the standard of living notably. (11) But the price was the silence of the grave for all criticism or opposition.

(12) El Benefactor's vanity grew with his personal wealth. (13) The jails were filled to overflowing with political prisoners who had incurred his displeasure. (14) He maintained amply financed lobbies in the United States and elsewhere which sycophantically chanted his praise, and his influence extended even to Congress.

(15) Until the last years or so the profession of profession of friendship with the United States had been an article of faith with Trujillo, and altogether too often this profession was accepted here as evidence of his good character. (16) Tardily the Government here came to understand how this country's own reputation was tarnished by the association with repression.

(17) Last year, after Trujillo had been cited for numerous aggressions in the Caribbean, the United States and many other members of the Organization of American States broke diplomatic relations with him.

(18) Thereupon followed a demonstration that tyranny knows no ideological confines. (19) Trujillo's dictatorship had been along conservative, right-wing lines. (20) But after the censure he and his propaganda started mouthing Communist slogans. (21) There was considerable evidence of a tacit rapprochement with Castro in Cuba, previously a *bête noire* to Trujillo - thus illustrating the way in which totalitarianism of the right and left coalesces.

(22) What comes after Trujillo is now the puzzle. (23) The Dominican people have known no democratic institutions and precious little freedom for a generation, and all alternative leadership has been suppressed. (24) Perhaps the army will be able to maintain stability, but the vacuum of free institutions creates a great danger. (25) The Dominican Republic could turn toward Communist-type authoritarianism as easily as toward Western freedom. (26) Such a twist would be a tragedy for the Dominican people, who deserve to breathe without fear. (27) For that reason any democratic reform and effort to bring genuine representative government to the Dominican Republic will need the greatest sympathy and help. (B01 128-180)

Segmented output:

(1) The end of Trujillo■

(2) Assassination■
even of a tyrant■
is repulsive to men of good conscience■

(3) Rafael Trujillo■
the often blood-thirsty dictator of the Dominican Republic for 31 years■
perhaps deserved his fate in an even-handed appraisal of history■

(5) Trujillo knew a great deal about assassination■

(6) The responsibility for scores of deaths■
including the abduction and murder of Jesus Maria Galindez■
a professor at Columbia University in New York■
has been laid at his door■

(7) He had been involved in countless schemes■
to do away with democratic leaders in neighboring countries■
such as President Romulo Betancourt of Venezuela■

(8) It was a sort of poetic justice■
that at the time of his own demise■
a new plot to overthrow the Venezuelan government■
reportedly involving the use of Dominican arms■
by former Venezuelan Dictator Marcos Perez Jimenez■

has been uncovered and quashed■
(9) The recent history of the Dominican Republic■
is an almost classical study of the way■
in which even a professedly benevolent dictatorship tends to become oppressive■
(10) Unquestionably■
Trujillo did some good things for his country■
he improved public facilities■
such as roads and sanitation■
attracted industry and investment■
and raised the standard of living notably■
(11) But the price was the silence of the grave■
for all criticism or opposition■
(12) El Benefactor's vanity grew with his personal wealth■
(13) The jails were filled to overflowing■
with political prisoners■
who had incurred his displeasure■
(14) He maintained amply financed lobbies■
in the United States and elsewhere■
which sycophantically chanted his praise■
and his influence extended even to Congress■
(15) Until the last years or so the profession of friendship with the United States■
had been an article of faith with Trujillo■
and altogether too often■
this profession was accepted here■
as evidence of his good character■
(16) Tardily the Government here came to understand■
how this country's own reputation was tarnished■
by the association with repression■
(17) Last year■
after Trujillo had been cited■
for numerous aggressions in the Caribbean■
the United States■
and and many other members of the Organization of American States■
broke diplomatic relations with him■
(18) Thereupon followed a demonstration■
that tyranny knows no ideological confines■
(19) Trujillo's dictatorship had been along conservative■
right-wing lines■
(20) But after the censure■

he and his propaganda started mouthing Communist slogans■

(21) There was considerable evidence of a tacit rapprochement■
with Castro in Cuba■
previously a *bête noire* to Trujillo■
thus illustrating the way in which totalitarianism of the right and left coalesces■

(22) What comes after Trujillo■
is now the puzzle■

(23) The Dominican people have known no democratic institutions■
and precious little freedom for a generation■
and all alternative leadership■
has been suppressed■

(24) Perhaps the army will be able to maintain stability■
but the vacuum of free institutions■
creates a great danger■

(25) The Dominican Republic could turn toward Communist-type authoritarianism■
as easily as toward Western freedom■

(26) Such a twist would be a tragedy■
for the Dominican people■
who deserve to breathe without fear■

(27) For that reason■
any democratic reform and effort■
to bring genuine representative government■
to the Dominican Republic■
will need the greatest sympathy and help■

# 13

# A Prolog implementation of automatic segmentation

*Mats Eeg-Olofsson*

## 13.1 Introduction

The TESS segmentation program successively reads sentences from an input text that has been tagged on the word and phrase levels (see Chapter 4). The input, which has the form of a labelled bracketing, is stored in a global 'chart' data structure, which can be accessed by the segmentation rules. The output of the segmentation procedure proper consists of lists of intervals, which describe the primary and secondary segments, referring to the rules that have been used to establish them. (See Chapter 12 for a detailed description of the rules and some output from the segmentation program.) Each interval description contains a pair of numbers, which denote the position in the input of the segment's first and last word, respectively. The interval lists can be used to retrieve the sequences of input words that belong to the respective segments.

Like the phrase level parser, the segmentation program is written in Sussex Prolog and runs on a VAX-11/730 computer under the VAX/VMS operating system. It takes about three hours to segment the test text (Brown Corpus B01) in this way. While this is too slow for interactive use, the Prolog implementation has important advantages for experimental work, owing to the great ease with which the segmentation rules can be changed. The following description is a sketch of a program that is but one in a series with varying designs. (Input arguments to the Prolog predicates are preceded by plus (+) signs, output arguments by minus (-) signs.) In the listing below, program code is indented; comments in the code are preceded by the percentage sign %; running text starts in the left-hand margin.

## 13.2 Top level predicates

The top segmentation predicate SEGMENTATION of the form

```
segmentation(+Start,+Stop,-Segs)
```

segments the part of the input between positions Start and Stop, producing a list (Segs) of primary segment descriptions. The descriptions are structures of the form

```
primary(rule(Rule), interval(Begin,End), secondary(Secsegs))
```

where Rule is the segmentation rule used to delimit the primary segment, Begin and End are positions in the input describing its boundaries, and Secsegs is a list of descriptions of the secondary segments contained in the primary segment.

SEGMENTATION uses PRIMARY_SEGMENT to get a primary segment starting in position Start, further SPLIT_PRIMARY to split up this primary segment into secondary segments, and then calls itself recursively to segment the remaining input.

```
segmentation(S,S,[]).
   % No input left - terminal case
segmentation(Start,Stop, [primary(rule(Rule),
   interval(Start,Break), secondary(Secsegs)) | Primsegs]) :-
   primary_segment(Start,Stop,Break,Rule),
   % Get a primary segment
   split_primary(Start,Break,Secsegs),
   % Split it up
   segmentation(Break,Stop,Primsegs).
   % Segment rest
```

PRIMARY_SEGMENT of the form

```
primary_segment(+Start,+Stop,-Break,-Rule)
```

is called by SEGMENTATION to find a primary segment boundary in position Break between Start and Stop. After skipping leading punctuation marks, PRIMARY_SEGMENT uses MAJOR_BREAK to find a boundary according to the rule Rule.

```
primary_segment(S,Stop,Break,Rule) :- psign(S), !,
   % Punctuation mark in position S
   S1 is S+1, primary_segment(S1,Stop,Break,Rule).
   % Skip punctuation mark
primary_segment(Tfirst,Tlast,Break,Rule) :-
   major_break(Tfirst,Tlast,Break,Rule).
   % Get real boundary
```

MAJOR_BREAK of the form `major_break(+First,+Last,-Break,-Rule)` tries to find a primary segment boundary in position Break between First and Last by applying a major segmentation rule Rule on some phrase boundary, checking the constraints (see Chapter 12, p 303). Otherwise, the whole remaining input between First and Last is taken to be the primary segment.

```
major_break(Tfirst,Tlast,Break,Rule) :-
   phrases_between(Tfirst,Break,Tlast),
   % Get phrase boundary
   not( general_constraints(Tfirst,Tlast,Break) ),
   % Check general constraints
   major_rule(Rule), % Get rule
   apply_rule(Rule,Tfirst,Tlast,Break), % Try rule
   not( constraints(Tfirst,Tlast,Break,Rule) ).
   % Check remaining constraints
major_break(Tfirst,Tlast,Tlast,last_primary) :-
   % Last primary segment
   last_in_input(Tlast).
```

SPLIT_PRIMARY, called by SEGMENTATION, of the form

```
split_primary(+From,+To,-Segs)
```

splits up the primary segment between positions From and To into a sequence of secondary segments, described by the list Segs. Each element of the list is a structure of the form `secondary(rule(Rule), interval(Begin,End))`, analogous to the descriptions of primary segments. SPLIT_PRIMARY skips punctuation marks, uses MINOR_BREAK to find some secondary segment boundary and then calls itself recursively to segment the rest of the primary segment.

```
split_primary(S,S,[]) :- !.
   % Nothing left -- terminal case
split_primary(S,Stop,Segs) :-
   psign(S), !, % Skip punctuation mark
   S1 is S+1, split_primary(S1,Stop,Segs).
split_primary(Tfirst,Tlast,
   [secondary(rule(Rule), interval(Tfirst,Break)) |
   Restsegs]) :-
   minor_break(Tfirst,Tlast,Break,Rule),
   % Find segment boundary by rule
   split_primary(Break,Tlast,Restsegs).
   % Find remaining secondary segments
```

```
split_primary(Tfirst,Tlast, % Last subsegment
   [secondary(rule(last_secondary),
   interval(Tfirst,Tlast))]).
```

MINOR_BREAK of the form

```
minor_break(+First,+Last,-Break,-Rule)
```

tries to find a secondary segment boundary in position Break between First and Last by applying a minor segmentation rule on some phrase boundary, checking the constraints.

```
minor_break(Tfirst,Tlast,Break,Rule) :-
   phrases_between(Tfirst,Break,Tlast),
   % Get phrase boundary
   not( general_constraints(Tfirst,Tlast,Break) ),
   % Check constraints
   minor_rule(Rule), % Get rule
   apply_rule(Rule,Tfirst,Tlast,Break), % Try rule
   not( constraints(Tfirst,Tlast,Break,Rule) ).
   % Check remaining constraints
```

## 13.3 Descriptions of rules

For convenience, the predicate RULEDESC offers a brief description of the segmentation rules to be implemented:

```
ruledesc(rule_1,
   'Punctuation replacement').
ruledesc(rule_2,
   'Before syndetic coordinator').
ruledesc(rule_3,
   'After initial and medial subordinate clauses').
ruledesc(rule_4,
   'Before medial and final adverbial clauses').
ruledesc(rule_5,
   'Before apposition markers').
ruledesc(rule_6,
   'Before nominal and relative clauses').
ruledesc(rule_7,
   'In postverbal adverbial/complement sequences').
ruledesc(rule_8,
   'After clause-initial adverbials').
ruledesc(rule_9,
   'After a complex subject').
ruledesc(rule_10,
   'After medial prepositional phrases').
ruledesc(rule_11,
   'In sequences of three prepositional phrases').
```

Major rules are used in the first cycle:

```
major_rule(rule_1).
major_rule(rule_2).
```

```
major_rule(rule_3).
major_rule(rule_4).
major_rule(rule_5).
```

The rest are minor rules, used in the second cycle:

```
minor_rule(Rule) :-
   ruledesc(Rule,_), not( major_rule(Rule) ).
```

Rule application proper is handled by the predicate APPLY_RULE, which has four arguments: a rule (rule_1, rule_2, ... ), the beginning position of the text segment to be subdivided by the rule, the end position of the text segment, and the breaking point, where the new tone unit boundary is to be introduced. As an illustration, the clauses corresponding to Rule 2 ('Before syndetic coordinator'), Rule 5 ('Before apposition markers'), Rule 9 ('After a complex subject'), and Rule 10 ('After medial prepositional phrases'; all documented in Chapter 12, pp 296ff) are included:

```
% Before syndetic coordinator:

% Clause for the case: of + X1 + CA + X2 + TO
apply_rule(rule_2, Tfirst, Tlast, Break) :-
   tag_at(Break,'CA'), Break < Tlast,
   Break > Tfirst, % CA
   analysed(B2,Break,word(X1)), B4 is B2-1,
   B4 >= Tfirst, word_at(B4,Of), word(Of,of),
   % of + X1
   B1 is Break+1, analysed(B1,B3,word(X2)), tag_at(B3,'TO').
   % X2 + TO

apply_rule(rule_2, Tfirst, Tlast, Break) :-
   tag_at(Break,Tag), Break < Tlast,
   Break > Tfirst,
   ( prefix(Tag,'CA') ; prefix(Tag,'CB') ;
   prefix(Tag,'CR') ),
   not( tight_coord(Break) ).
   % Conditions a) - f) must not apply

% Before apposition markers:

apply_rule(rule_5, Tfirst, Tlast, Break) :-
   word_at(Break,App), has_tag(App,AppT),
   prefix(AppT,'ACapp'),
   Break < Tlast, Break > Tfirst, word(App,AppW),
   member(AppW,
   % Standard predicate for list membership
   ['for example','for instance',ie,'i.e.',eg,'e.g.','in
   other words',namely,'such as','that is']).
```

```
% Cycle 2
% After a complex subject:

apply_rule(rule_9, Tfirst, Tlast, Break) :-
   analysed(B0, Break, nph(_,_)), B0 >= Tfirst,
   Break > Tfirst, Break < Tlast,
   words_between(B0,Break,N), N > 2,
   % Condition b)
   analysed(Break, B1, vph(_,Vfeats)),

   ( words_between(Tfirst, Break, D), D > 4 ;
   % Condition a)
   chars_between(Tfirst, Break, C), C > 30 ),

   finite_feats(Vfeats),
   not( (member(pass,Vfeats),
        analysed(B1,_,pph(_,_)) ) ).

% After medial prepositional phrases

apply_rule(rule_10, Tfirst, Tlast, Break) :-
   analysed(B0,Break,pph(_,_)), B0 >= Tfirst,
   Break < Tlast, Break > Tfirst,
   analysed(Break,_,vph(_,Vfeats)), finite_feats(Vfeats).

% The following clauses for TIGHT_COORD each
% describe one of the conditions
% a) - f) for Rule 2 ('Before a coordinator'):

tight_coord(Conjpos) :- % Condition a)
   B0 is Conjpos-1, tag_at(B0,T1), maintag(T1,MainTag),
   B1 is Conjpos+1, tag_at(B1,T2), maintag(T2,MainTag).
   % Main parts of word class tags are identical

tight_coord(Conjpos) :- % Condition b)
   B0 is Conjpos-1, tag_at(B0,'BHsub').

tight_coord(Conjpos) :- % Condition c)
   word_at(Conjpos,Or), word(Or,or), has_tag(Or,'CR'),
   B1 is Conjpos+1, word_at(B1,X2),
   ( word(X2,so) ; word(X2,not) ). % or so/not

tight_coord(Conjpos) :- % Condition d)
   word_at(Conjpos,And), word(And,and), has_tag(And,'CA'),
   B1 is Conjpos+1, word_at(B1,EW),
   ( word(EW,elsewhere) ; word(EW,more) ).
   % and elsewhere/more

tight_coord(Conjpos) :- % Condition e)
   tag_at(Conjpos,CA), prefix(CA,'CA'),
   one_phrase(B1,Conjpos), B0 is B1-1, % X1
   word_at(B0,Betw), word(Betw,between). % between

tight_coord(Conjpos) :- % Condition f)
   tag_at(Conjpos,CT), prefix(CT,'CA'),
   C1 is Conjpos+1, word_at(C1,W1),
   ( word(W1,'I') ; has_tag(W1,'BHobj') ),
   C0 is Conjpos-1, tag_at(C0,T0),
   ( prefix(T0,'BHneg') ; prefix(T0,'BHper') ).

% Auxiliary predicate: Feature description of
% finite verb phrases:
```

```
finite_feats(Vfeats) :-
   not( member(ing,Vfeats) ),
   not( member(inf,Vfeats) ).
```

## 13.4 Constraints

Several general types of constraints are built into the segmentation procedure: the restriction that segment boundaries must coincide with phrase boundaries, encoded in the predicate PHRASES_BETWEEN, the restriction that contractions (e.g. *don't*) must not be split up, encoded by the predicate GENERAL_CONSTRAINTS, and a set of more specific constraints described by the predicate CONSTRDESC:

```
% General constraints: Do not split contractions

general_constraints(Tfirst,Tlast,Break) :-
   apc(contraction,Tfirst,Tlast,Break).

% Brief description of constraints (see Chapter 12, p 304
% for details):

constrdesc(constraint_c1,
   'Pronoun-verb sequences').
constrdesc(constraint_c2,
   'After certain light elements').
constrdesc(constraint_c3,
   'Before certain light adverbs').
constrdesc(constraint_c4,
   'After nonfinite verbs').
constrdesc(constraint_c5,
   'After PA in prepositional ing-clauses').
constrdesc(constraint_c6,
   'Certain indefinite pronouns + BUT').
```

The top-level constraint predicate CONSTRAINTS calls the auxiliary predicate APPLY_CONSTRAINT, which, in turn, calls APC to check the constraint, and DISPLAY_CONSTRAINT_SUCCESS (not documented here) to record its effect for later inspection.

```
constraints(Tfirst,Tlast,Break,Rule) :-
   constrdesc(C,_), % Get a constraint
   apply_constraint(C,Tfirst,Tlast,Break,Rule).
   % Check applicability

apply_constraint(Constraint,Tfirst,Tlast,Break,Rule) :-
   apc(Constraint,Tfirst,Tlast,Break), !,
   % Check constraint
   display_constraint_success(Constraint,
      Tfirst,Tlast,Break,Rule).
   % Record
```

The following clauses for the predicate APC, which describes constraint application proper, are included to illustrate the implementation of the constraint on contractions as well as two other more special constraints:

```
% Contractions:
apc(contraction, Tfirst, Tlast, Break) :-
   B1 is Break-1, contraction_at(B1).

% Pronoun-verb sequences
apc(constraint_c1, Tfirst, Tlast, Break) :-
   analysed(Break,_,word(Verb)), has_tag(Verb,Vtag),
   prefix(Vtag,'V'),
   analysed(_,Break,word(Preverb)), has_tag(Preverb,Pvtag),
   ( prefix(Pvtag,'AB') ; prefix(Pvtag,'BH'),
   Pvtag \= 'BHobj' ).

% After nonfinite verb phrases:
apc(constraint_c4, Tfirst, Tlast, Break) :-
   not( psign(Break) ),
   analysed(_,Break,vph(_,Vfeats)),
   not( finite_feats(Vfeats) ).
```

## 13.5 Access to input in chart

The segmentation predicates use the following routines to access the input in the chart, whose contents are described by the three predicates ANALYSED, LAST_IN_INPUT, and CHART_INDEX. ANALYSED is a collection of facts of the form `analysed(Vertex1,Vertex2,Phrase)`, where the first two arguments are positions in the input and the third is a structural description of a phrase. Structural descriptions are either of the form `Name(Daughters,Features)`, where Name is the name of any phrase type, or of the form `other(Word)` (see Section 4.3). Single words (including 'ditto-tagged' multi-word units) are represented as facts of the form `analysed(Vertex1,Vertex2,word(Word))`. The vertex (position) numbers range from 0 to the value stored in the only clause for the one-place predicate LAST_IN_INPUT. (The chart can be thought of as a graph with numbered vertices and edges labelled with phrases and words. Thus the interpretation of a fact such as `analysed(Vertex1, Vertex2,Phrase)` is that the vertices Vertex1 and Vertex2 are incident with an edge labelled Phrase.) The predicate CHART_INDEX of the form `chart_index(Vertex,Words)` is used to record the number of graphic words Words from the beginning of the input to the position Vertex. (Because of the existence of contractions and multi-word units, this number need not equal the number of tagged grammatical 'word' units, described by Vertex.)

PHRASES_BETWEEN of the form

```
phrases_between(+From,-To,+Before)
```

unifies To with some position of a phrase boundary between From and Before in the input. It is used to find phrase boundaries, which are the only candidates for breaking points.

```
phrases_between(From,To,Before) :-
    one_phrase(From,To), To < Before.
phrases_between(From,To,Before) :-
    one_phrase(From,X), X < Before,
    phrases_between(X,To,Before).

% One 'phrase' input between positions From and To
one_phrase(From,To) :- analysed(From,To,Phrase),
    Phrase =.. [Type,Dauts,Feats]. % Phrase proper
one_phrase(From,To) :- analysed(From,To,other(_)).
    % Other material

% Get word at position Ind in chart
word_at(Ind,Word) :-
    analysed(Ind,I1,word(Word)), I1 is Ind+1.

% Get tag at position Ind in chart
tag_at(Ind,Tag) :-
    word_at(Ind,Word), has_tag(Word,Tag).

% Punctuation mark in input in position Vertex
psign(Vertex) :-
    analysed(Vertex,V1,word(W)), V1 is Vertex+1,
    word(W,Word), punctuation_mark(Word).

% Count characters in input between positions
% Vertex1 and Vertex2
chars_between(Vertex1, Vertex2, Chars) :-
    word_sequence(Vertex1, Vertex2, Words),
    mapcar(stringlen,Words,Lengths),
    sum(Lengths,Length), length(Lengths,L),
    Chars is L+Length.

% Count graphic words in input between positions
% Vertex1 and Vertex2
words_between(Vertex1,Vertex2,Dist) :-
    chart_index(Vertex1,Ind1),
    chart_index(Vertex2,Ind2),
    Dist is Ind2-Ind1.

% Find contraction in position From
contraction_at(From) :-
    word_at(From,W1), word(W1,WP1), has_tag(W1,Tag1),
    % Part 1
    F1 is From+1,
    word_at(F1,W2), word(W2,WP2), has_tag(W2,Tag2),
    % Part 2
    posscontr(WP1,Tag1,WP2,Tag2).
    % Possible contraction

% Access to word sequence in input from position
% From to position To
word_sequence(From,From,[]) :- !.
word_sequence(From,To,[Word1|Rest]) :-
    word_at(From,W), word(W,Word1),
    F1 is From+1, word_sequence(F1,To,Rest).
```

## 13.6 Input routines

The input routines store analysed sentences in the chart. The input sentences are lists of labelled bracketings (see Section 4.3 for details), that are read as Prolog terms and encoded by the chart predicates ANALYSED, LAST_IN_INPUT, and CHART_INDEX (see Section 13.5).

```
% Load chart with input
load_chart(Analysis,End) :-
    erase_chart, % Erase old chart
    load_part(Analysis,0,End),
    % Store phrases and words
    asserta( last_in_input(End) ),
    % Record last position in input
    index_chart(End).
    % Count distances in graphic words

% Erase information in old chart
erase_chart :-
    retractall(analysed(_,_,_)),
    % Delete all clauses for ANALYSED (standard)
    retractall(last_in_input(_)),
    retractall(chart_index(_,_)).

% Load list of input items (phrases or 'other')
load_part([],X,X) :- !.
    % Nothing left -- terminal case
load_part([Part1|Rest],Start,Stop) :-
    load_phrase(Part1,Start,X),
    % Store phrase and words of first item
    load_part(Rest,X,Stop).
    % Load remaining items


% Load input item
load_phrase(other(X),Start,Stop) :- !,
    % 'Other' input item
    Stop is Start+1,
    lower_brown_word(X,Y),
    % Get word and convert to lower case
    asserta( analysed(Start,Stop,other(word(Y))) ),
    % Store as 'other'
    asserta( analysed(Start,Stop,word(Y)) ).
    % Store as word
load_phrase(P,Start,Stop) :-
    P =.. [Type,Dauts,Feats],
    % Phrase proper
    words(Dauts,Ws), length(Ws,Wl),
    Stop is Start+Wl,
    asserta( analysed(Start,Stop,P) ),
    % Store phrase
    load_words(Ws,Start).
    % Store component words

% Store an input word list
load_words([],_) :- !.
    % No words left -- terminal case
```

```
load_words([W1|Wrest],S) :-
    S1 is S+1,
    asserta( analysed(S,S1,W1) ), % Store word
    load_words(Wrest,S1). % Store remaining words

% Get component words in a phrase
words(Dauts,Words) :- ws(Dauts,Words,[]).

% Auxiliary predicate WS finds the component words
% in a phrase by 'unpacking' its subconstituent
% descriptions ('daughters')
ws([],W,W) :- !.
ws(word(W),New,Old) :- !,
    lower_brown_word(W,L), % Convert to lower case
    append(Old,[word(L)],New).
ws(hole(nil),Old,Old) :- !. % Empty hole
ws(hole(X),New,Old) :- !,
    ws(X,New,Old).
ws(P,New,Old) :-
    P =.. [Type,Dauts,Feats], Type \= '.', !,
    ws(Dauts,New,Old).
ws([I1|Items],New,Old) :-
    ws(I1,N1,[]),
    append(Old,N1,N2),
    ws(Items,New,N2).

% Compute and store values of function CHART_INDEX
% from right to left
index_chart(0) :- !,
    % Beginning of input -- terminal case
    assertz( chart_index(0,0) ).
index_chart(Last) :-
    L2 is Last-2, contraction_at(L2), !,
    % Contraction
    index_chart(L2), chart_step(L2,S2),
    L1 is Last-1, chart_step(L1,S1),
    chart_index(L2,Ind2), Ind is Ind2+S2+S1-1,
    assertz( chart_index(Last,Ind) ).
index_chart(Last) :- % No contraction
    L1 is Last-1, index_chart(L1),
    chart_index(L1,Ind), chart_step(L1,Step),
    Ind1 is Ind+Step,
    assertz( chart_index(Last,Ind1) ).

% Number of graphic words in input from position
% Ind to Ind+1
chart_step(Ind,0) :- psign(Ind), !.
    % Punctuation marks excluded
chart_step(Ind,Step) :-
    word_at(Ind,Word),
    has_tag(Word,Tag), tagval(Tag,Step).
```

## 13.7 Auxiliary predicates

The segmentation program also uses the following predicates, whose definitions will not be specified in detail:

Access to word class tag structure:

| | |
|---|---|
| POSSCONTR | test for contraction |
| TAGVAL | access to number of ditto-tagged words |
| MAINTAG | access to main part of tag (excluding morphology) |

Punctuation marks:

| | |
|---|---|
| PUNCTUATION_MARK | check if 'word' is a punctuation mark |

Access to tagged Brown corpus input in 'vertical' format:

| | |
|---|---|
| HAS_TAG | access to full word class tag (including length) |
| WORD | access to word body |
| LOWER_BROWN_WORD | convert word body to lower case |

String handling:

| | |
|---|---|
| PREFIX | test for prefix |
| STRINGLEN | length of string |

List handling: (standard)

| | |
|---|---|
| APPEND | list concatenation |
| MEMBER | list membership |
| LENGTH | list length |

Miscellaneous:

| | |
|---|---|
| MAPCAR | list mapping (cf LISP functional) |
| SUM | add up a list of numbers |

# References

This list includes works referred to in the body of text in this book. There is also a bibliography in Chapter 1, Appendix 2, which includes publications that have used Survey of English Usage material.


*A Manual of Style*. 1969. Chicago: The University of Chicago Press.

Aarts, J. & W. Meijs (eds). 1984. *Corpus linguistics: Recent developments in the use of computer corpora in English language research.* Amsterdam: Rodopi.

Aarts, J. & W. Meijs (eds). 1986. *Corpus linguistics II: New studies in the analysis and exploitation of computer corpora.* Amsterdam: Rodopi.

Aijmer, K. 1984. *Sort of* and *kind of* in English conversation. *Studia Linguistica* 38, 118-128.

Aijmer, K. 1986. Discourse variation and hedging. In Aarts & Meijs 1986, 1-18.

Aijmer, K. 1987. *Oh* and *ah* in English conversation. In Meijs 1987, 61-86.

Allen, J. 1985a. Scanning the issue. The Special Issue on Man-Machine Speech Communication. *Proceedings of the IEEE* 73, 1539-1540.

Allen, J. 1985b. A perspective on man-machine communication by speech. The Special Issue on Man-Machine Speech Communication. *Proceedings of the IEEE* 73, 1541-1550.

Allen, J., M.S. Hunnicut & D. Klatt. 1987. *From text to speech: the MITalk system.* Cambridge Studies in Speech Science and Communication. Cambridge: Cambridge University Press.

Allerton, D.J. & A. Cruttenden. 1974. English sentence adverbials: Their syntax and their intonation in British English. *Lingua* 34, 1-30.

Allerton, D.J. & A. Cruttenden. 1976. The intonation of medial and final sentence adverbials in British English. *Archivum Linguisticum* 7, 29-59.

Allerton, D.J. & A. Cruttenden. 1978. Syntactic, illocutionary, thematic and attitudinal factors in the intonation of adverbials. *Journal of Pragmatics* 2, 155-188.

Altenberg, B. 1986. ICAME bibliography. *ICAME News* 10, 62-79.

Altenberg, B. 1987a. *Prosodic patterns in spoken English. Studies in the correlation between prosody and grammar for text-to-speech conversion.* Lund Studies in English 76. Lund: Lund University Press.

Altenberg, B. 1987b. Predicting text segmentation into tone units. In Meijs 1987, 49-60.

Altenberg, B. 1987c. Some functions of the booster in spoken English. In Lindblad & Ljung 1987, 5-20.

Altenberg, B. 1987d. Causal ordering strategies in English conversation. In Monaghan 1987, 50-64.

Altenberg, B. 1988. Spoken English and the lexicon. In Liedloff 1988, 19-32.

Altenberg, B. Forthcoming a. ICAME Bibliography. In Johansson & Stenström (forthcoming).

Altenberg, B. Forthcoming b. Amplifier collocations in spoken English. In Johansson & Stenström (forthcoming).

Altenberg, B. Forthcoming c. Speech as linear composition. To appear in a conference volume from the Fourth Nordic Conference for English Studies (Helsingör, 11 May 1989).

Altenberg, B. & M. Eeg-Olofsson. Forthcoming. Phraseology in spoken English: Presentation of a project. To appear in a volume in the series *Language and computers: Studies in Practical Linguistics*, eds J. Aarts & W. Meijs. Amsterdam: Rodopi.

Andersson, U., T. Lindgren & F. Jungermann. 1987. CLINT. Segmenteringsprogram för engelsk text. Användarmanual och kodbeskrivning. Lund University: Department of Computer Sciences.

Atkinson, J. & J. Heritage (eds). 1984. *Structures of social action.* Cambridge: Cambridge University Press.

Bäckman, S. & G. Kjellmer (eds). 1985. *Papers on language and literature presented to Alvar Ellegård and Erik Frykman.* Gothenburg Studies in English 60. Göteborg: Acta Universitatis Gothoburgensis.

Barik, H.C. 1968. On defining juncture pauses: a note on Boomer's 'Hesitation and grammatical encoding'. *Language and Speech* 11, 156-159.

Beattie, G. 1983. *Talk.* Milton Keynes: Open University Press.

Biber, D. 1986a. Spoken and written textual dimensions in English: Resolving the contradictory findings. *Language* 62, 384-414.

Biber, D. 1986b. On the investigation of spoken/written differences. *Studia Linguistica* 40, 1-21.

Biber, D. 1988. *Variation across speech and writing*. Cambridge: Cambridge University Press.

Bing, J. 1984. A discourse domain identified by intonation. In Gibbon & Richter 1984, 10-19.

Blankenship, J. & C. Kay. 1964. Hesitation phenomena in English Speech: A study in distribution. *Word* 20, 360-372.

Bolinger, D. 1972a. *Degree words*. The Hague: Mouton.

Bolinger, D. 1972b. Accent is predictable (if you're a mindreader). *Language* 48, 633-644.

Bolinger, D. 1986. *Intonation and its parts. Melody in spoken English.* London: Edward Arnold.

Bond, Z. S. & S. Garner. 1980. Misperceptions of fluent speech. In Cole 1980, 115-132.

Boomer, D. 1965. Hesitation and grammatical encoding. *Language and Speech* 8, 148-158.

Boomer, D. 1970. Review of Goldman-Eisler 1968. *Lingua* 25, 152-164.

Boomer, D. & J. Laver. 1968. Slips of the tongue. *British Journal of Disorders of Communication* 3, 2-12.

Brotherton, P. 1979. Speaking and not speaking: Process for translating ideas into speech. In Siegman & Feldstein 1979, 179-209.

Brown, G. 1977. *Listening to spoken English.* London: Longman.

Brown, G., K. Currie & J. Kenworthy. 1980. *Questions of intonation.* London: Croom Helm.

Butterworth, B. 1975. Hesitation and semantic planning in speech. *Journal of Psycholinguistic Research* 4, 75-87.

Butterworth, B. (ed). 1980a. *Language production,* vol 1: *Speech and talk.* London: Academic Press.

Butterworth, B. 1980b. Evidence from pauses in speech. In Butterworth 1980a, 155-176.

Carlson, L. 1984. *'Well' in dialogue games. A discourse analysis of the interjection 'well' in idealized conversation.* Amsterdam: John Benjamins.

Carter, R. & M. McCarthy (eds). 1988. *Vocabulary and language teaching.* London: Longman.

Chafe, W. 1979. The flow of thought and the flow of language. In Givón 1979, 159-181.

Chafe, W. (ed). 1980a. *The pear stories.* Norwood, N.J.: Ablex.

Chafe, W. 1980b. The deployment of consciousness in the production of a narrative. In Chafe 1980a, 9-50.

Chafe, W. 1987. Cognitive constraints on information flow. *Coherence and grounding in discourse*, ed R.S. Tomlin, 21-51. Amsterdam: John Benjamins.

Chafe, W. & J. Danielewicz. 1987. Properties of spoken and written language. In *Comprehending oral and written language*, ed R. Horowitz & S.J. Samuels, 83-113. New York: Harcourt Brace Jovanovich.

Clark, H.H. & E.V. Clark. 1977. *Psychology and language. An introduction to psycholinguistics.* New York: Harcourt Brace Jovanovich.

Coker, C., N. Umeda & P. Browman. 1973. Automatic synthesis from ordinary English text, *IEEE Transactions on audio and electroacoustics*, Au-21, 293-298.

Cole, R.A. (ed). 1980. *Perception and production of fluent speech.* Hillsdale, N.J.: Lawrence Erlbaum.

Cole, R.A. & J. Jakimik. 1980. A model of speech perception. In Cole 1980, 133-163.

*Collins COBUILD English language dictionary.* 1987. London: Collins.

Cook, M. 1971. The incidence of filled pauses in relation to part of speech. *Language and Speech* 14, 135-139.

Coulthard, M. 1981. Developing the description. In Coulthard & Montgomery 1981, 13-30.

Coulthard, M. & M. Montgomery (eds). 1981. *Studies in discourse analysis.* London: Routledge & Kegan Paul.

Cruttenden, A. 1986. *Intonation*. Cambridge: Cambridge University Press.

Crystal, D. 1967. English. *Word classes*, Volume reprinted from *Lingua* 17, 24-56. Amsterdam: North-Holland.

Crystal, D. 1969. *Prosodic systems and intonation in English.* Cambridge: Cambridge University Press.

Crystal, D. 1975. *The English tone of voice. Essays on intonation, prosody and paralanguage.* London: Edward Arnold.

Crystal, D. 1980. Neglected grammatical factors in conversational English. In Greenbaum, Leech & Svartvik 1980, 153-166.

Crystal, D. & D. Davy. 1969. *Investigating English style.* London: Longman.

Crystal, D. & D. Davy. 1975. *Advanced conversational English.* London: Longman.

Crystal, D. & R. Quirk. 1964. *Systems of prosodic and paralinguistic features in English.* The Hague: Mouton.

Dechert, H.W. & M. Raupach (eds). 1980. *Temporal variables in speech. Studies in honour of Frieda Goldman-Eisler.* The Hague: Mouton.

Déjean le Féal, K. 1982. Why impromptu speech is easy to understand. In Enkvist 1982, 221-239.

Du Bois, J.W., S. Schuetze-Coburn & S. Cumming. 1989. *Discourse Transcription. A Handbook.* University of California, Santa Barbara.

Edmondson, W. 1981. *Spoken discourse.* London: Longman.

Edwards, J.A. 1989. *Transcription and the new functionalism: A counterproposal to CHILDES' CHAT conventions.* Berkeley Cognitive Science Report No. 58, March 1989. Institute of Cognitive Studies, University of California, Berkeley.

Eeg-Olofsson, M. 1985. A probability model for computer-aided word-class determination. *ALLC Journal* 5, 25-30.

Eeg-Olofsson, M. 1987. Assigning new tags to old texts - an experiment in automatic word class tagging. In Meijs 1987, 45-47.

Eeg-Olofsson, M. & J. Svartvik. 1984. Four-level tagging of spoken English. In Aarts & Meijs 1984, 53-64.

Enkvist, N.E. (ed). 1982. *Impromptu speech: A symposium.* Åbo: Åbo Akademi.

Erman, B. 1987. *Pragmatic expressions in English: A study of 'you know', 'you see' and 'I mean' in face-to-face conversation.* Stockholm Studies in English 69. Stockholm: Almqvist & Wiksell.

Fallside, F. & Woods, W.A. (eds) 1985. *Computer speech processing.* Englewood Cliffs, N.J.: Prentice-Hall.

Fjelkestam-Nilsson, B. 1983. *'Also' and 'too': A corpus-based study of their frequency and use in modern English.* Stockholm Studies in English 58. Stockholm: Almqvist & Wiksell.

Francis, W.N. 1980. A tagged corpus - problems and prospects. In Greenbaum, Leech & Svartvik 1980, 192-209.

Francis, W.N. & H. Kučera. 1982. *Frequency analysis of English usage: Lexicon and grammar*. Boston: Houghton Mifflin Company.

Freedle, R.O. (ed). 1977. *Discourse production and comprehension*. Vol I, Discourse Processes: Advances in Research and Theory. Norwood, N.J.: Ablex.

Freedle, R.O. (ed) 1979. *New directions in discourse processing*. Vol II, Advances in Discourse Processes. Norwood, N.J.: Ablex.

Gårding, E. 1967. Prosodiska drag i spontant och uppläst tal. In G. Holm (ed), *Svenskt talspråk*, 40-85. Stockholm: Almqvist & Wiksell.

Garside, R., G. Leech & G. Sampson (eds). 1987. *The computational analysis of English. A corpus-based approach*. London: Longman.

Gibbon, D. & H. Richter (eds). 1984. *Intonation, accent and rhythm. Studies in discourse phonology*. Berlin: Walter de Gruyter.

Givón, T. (ed). 1979. *Syntax and semantics, 12: Discourse and syntax*. New York: Academic Press.

Goldman-Eisler, F. 1968. *Psycholinguistics: Experiments in spontaneous speech*. London: Academic Press.

Goldman-Eisler, F. 1980. Psychological mechanisms of speech production as studied through the analysis of simultaneous translation. In Butterworth 1980, 143-153.

Greenbaum, S. 1969. *Studies in English adverbial usage*. London: Longman.

Greenbaum, S. 1988. *Good English and the grammarian*. London: Longman.

Greenbaum, S., G. Leech & J. Svartvik (eds). 1980. *Studies in English linguistics for Randolph Quirk*. London: Longman.

Grosjean, F. 1980. Linguistic structures and performance structures: Studies in pause distribution. In Dechert & Raupach 1980, 91-106.

Halliday, M.A.K. 1967. *Intonation and grammar in British English*. The Hague: Mouton.

Halliday, M.A.K. 1985a. *Spoken and written language*. Deakin: Deakin University Press.

Halliday, M.A.K. 1985b. *An introduction to functional grammar*. London: Edward Arnold.

Halvorsen, P-K. 1988. Computer applications in linguistic theory. *Linguistics: The Cambridge Survey*, vol II, *Linguistic theory: Extensions and implications*, ed. F.J.Newmeyer, 198-219. Cambridge: Cambridge University Press.

Henderson, A. 1974. Time patterns in spontaneous speech - cognitive stride or random walk? A reply to Jaffe et al. *Language and Speech* 17, 119-125.

Henderson, A., F. Goldman-Eisler & A. Skarbeck. 1966. Sequential temporal patterns in spontaneous speech. *Language and Speech* 9, 207-216.

Heritage, J. 1984. A change-of state token and aspects of its sequential place. In Atkinson & Heritage 1984, 299-345.

Hofland, K. & S. Johansson. 1982. *Word frequencies in British and American English.* Bergen: The Norwegian Computing Centre for the Humanities.

Householder, F.W. 1982. Analysis, synthesis and improvisation. In *Text processing*, ed S. Allén, 271-295. Stockholm: Almqvist & Wiksell.

Ilson, R. (ed). 1985. *Dictionaries, lexicography and language learning.* Oxford: Pergamon Press.

Isaksson, L., A. Landin, O. Persson, K. Tsilkos & M. Wilson. 1986. PROSODY - ett program avsett för systematisk grafisk framställning av engelskt talspråk. Lund University: Department of Computer Science.

Jacobson, S. 1978. *On the use, meaning, and syntax of English preverbal adverbs.* Stockholm Studies in English 44. Stockholm: Almqvist & Wiksell.

James, D. 1972. Some aspects of the syntax and semantics of interjections. *Papers from the 8th regional meeting of the Chicago Linguistic Society*, 162-172.

Johansson, S. (ed). 1982. *Computer corpora in English language research.* Bergen: Norwegian Computing Centre for the Humanities.

Johansson, S. 1986. *The tagged LOB corpus. Users' manual.* Bergen: Norwegian Computing Centre for the Humanities.

Johansson, S. 1987. Machine-readable texts in English language research: Progress and prospects. In Lindblad & Ljung 1987, 125-137.

Johansson, S. & K. Hofland. 1989. *Frequency analysis of English vocabulary and grammar*, volumes 1 & 2. Oxford: Oxford University Press.

Johansson, S. & M-C. Jahr. 1982. Grammatical tagging of the LOB Corpus: Predicting word class from word endings. In Johansson 1982, 118-146.

Johansson, S. & A-B. Stenström (eds). Forthcoming. *English computer corpora.* Topics in English Linguistics. Berlin: Mouton de Gruyter.

Johns-Lewis, C. (ed). 1986. *Intonation in discourse.* London: Croom Helm.

Jönsson, G., B. Mandersson & B. Sigurd. 1982. A microcomputer pausemeter for linguists. Working Papers 24, 1-20. Lund University: Department of Linguistics.

Kasl, S. & G. Mahl. 1965. The relationship of disturbances and hesitations in spontaneous speech to anxiety. *Journal of Personality and Social Psychology* 1, 425-433.

Knowles, G. 1987. *Patterns of spoken English: An introduction to English phonetics.* London: Longman.

Knowles, G. & L. Lawrence. 1987. Automatic intonation assignment. In Garside, Leech & Sampson 1987, 139-148.

Kučera, H. & W. N. Francis. 1967. *Computational analysis of present-day American English.* Providence, R.I.: Brown University Press.

Kühlwein, W. & B. Spillner (eds). 1988. *Sprache und Individuum.* Tübingen: Gunter Narr.

Kytö, M. & O. Ihalainen & M. Rissanen (eds). 1988. *Corpus linguistics, hard and soft.* Amsterdam: Rodopi.

Labov, W. & D. Fanshel. 1977. *Therapeutic discourse.* New York: Academic Press.

Lakoff, R. 1973. Questionable answers and answerable questions. *Issues in linguistics. Papers in honor of Henry and Renee Kahane*, ed B. Kachru, 453-467. Urbana: University of Illinois Press.

Lallgee, M. & M. Cook. 1969. An experimental investigation of the function of filled pauses in speech. *Language and Speech* 12, 24-28.

Leech, G. 1987. General introduction. In Garside, Leech & Sampson 1987, 1-15.

Leech, G., R. Garside & E.S. Atwell. 1983. Recent developments in the use of computer corpora in English language research. *Transactions of the Philological Society*, 32-40.

Leech, G., R. Garside & E.S. Atwell. 1983. The automatic grammatical tagging of the LOB Corpus. *ICAME News* 7, 13-33.

Leech, G. & J. Svartvik. 1975. *A communicative grammar of English.* London: Longman.

Lieberman, P. 1986. The acquisition of intonation by infants: Physiology and neural control. In Johns-Lewis 1986, 239-257.

Lieberman, P. & S.E. Blumstein. 1988. *Speech physiology, speech perception, and acoustic phonetics.* Cambridge Studies in Speech Science and Communication. Cambridge: Cambridge University Press.

Liedloff, V. (ed). 1988. *Standardization in computerized lexicography. Proceedings of the IV Forum for Information Science and Practice.* Saarbrücken: IAI.

Lindblad, I. & M. Ljung (eds). 1987. *Proceedings from the Third Nordic Conference for English Studies*, vol I. Stockholm Studies in English 73. Stockholm: Almqvist & Wiksell.

Lindquist, H. 1989. *English adverbials in translation: A corpus study of Swedish renderings.* Lund Studies in English 80. Lund: Lund University Press.

Little, L. 1963. *The syntax of vocalized pauses in spontaneous cultivated speech.* Ann Arbor, Michigan: University Microfilms International.

*Longman dictionary of contemporary English.* 1987. 2nd ed. London: Longman.

Lounsbury, F. 1954. Transitional probability, linguistic structure, and systems of habit-family hierarchies. *Journal of Abnormal and Social Psychology* 49, 93-101.

Maclay, H. & C. Osgood. 1959. Hesitation phenomena in spontaneous English speech. *Word* 15, 19-44.

Marshall, I. 1983. Choice of grammatical word-class without global syntactic analysis: Tagging words in the LOB Corpus. *Computers and the Humanities* 17, 139-150.

Meijs, W. (ed). 1987. *Corpus linguistics and beyond.* Amsterdam: Rodopi.

Merrit, M. 1984. On the use of 'okay' in service encounters. *Language in use*, ed J. Baugh & J. Sherzer, 139-147. Englewood Cliffs, N.J.: Prentice-Hall.

Meyer, C.F. 1986. Punctuation practice in the Brown Corpus. *ICAME News* 10, 80-95.

Monaghan, J. (ed). 1987. *Grammar in the construction of texts*. London: Frances Pinter.

Nässlin, S. 1984. *The English tag question: A study of sentences containing tags of the type 'isn't it?', 'is it?'*. Stockholm Studies in English 60. Stockholm: Almqvist & Wiksell.

O'Connell, D. & S. Kowal. 1983. Pausology. *Computers in language research* 2, 221-301. Berlin: Walter de Gruyter.

Oreström, B. 1983. *Turn-taking in English conversation*. Lund Studies in English 66. Lund: Lund University Press.

Östman, J-O. 1981. *'You know': a discourse-functional approach*. Amsterdam: John Benjamins.

*Oxford dictionary of current idiomatic English. Volume 2: phrase, clause & sentence idioms*, eds A.P. Cowie, R. Mackin & I.R. McCaig. 1983. Oxford: Oxford University Press.

Pisoni, D.B., H.C. Nusbaum & B.G. Greene. 1985. Perception of synthetic speech generated by rule. *Proceedings of the IEEE* 73, 1665-1676.

Quirk, R., A. Duckworth, J. Rusiecki, J. Svartvik & A. Colin. 1964. Studies in the correspondence of prosodic to grammatical features in English. *Proceedings of the Ninth International Congress of Linguists*, 679-691. The Hague: Mouton. [Reprinted in R. Quirk, *Essays on the English language: Medieval and modern*, 120-135. London: Longman, 1968.]

Quirk, R., S. Greenbaum, G. Leech & J. Svartvik. 1985. *A comprehensive grammar of the English language*. London: Longman.

Renouf, A. 1987. Corpus development. In Sinclair 1987a, 1-40.

Ringbom, H. & M. Rissanen (eds). 1984. *Proceedings from the Second Nordic Conference for English Studies*. Åbo: Åbo Akademi.

Rochester, S. 1973. The significance of pauses in spontaneous speech. *Journal of psycholinguistic research* 2, 51-81.

Sacks, H., E. Schegloff & G. Jefferson. 1974. A simplest systematics for the organization of turn-taking for conversation. *Language* 50, 696-735.

Sahlin, E. 1979. *'Some' and 'any' in spoken and written English*. Studia Anglistica Upsaliensia 38. Stockholm: Almqvist & Wiksell.

Sampson, G. 1987a. Probabilistic models of analysis. In Garside, Leech & Sampson 1987, 16-29.

Sampson, G. 1987b. Alternative grammatical coding systems. In Garside, Leech & Sampson 1987, 165-183.

Schegloff, E. 1982. Discourse as an interactional achievement. Some uses of *uhuh* and other things that come between sentences. In Tannen 1982, 71-93.

Schegloff, E. & H. Sacks. 1973. Opening up closings. *Semiotica* 8, 289-327.

Schiffrin, D. 1986. *Discourse markers*. Cambridge: Cambridge University Press.

Schourup, L. 1985. *Common discourse particles in English conversation*. New York: Garland Publishing.

Siegman, A.W. & S. Feldstein (eds). 1979. *Of speech and time. Temporal speech patterns in interpersonal contexts*. Hillsdale, N.J.: Laurence Erlbaum.

Sigurd, B. 1983. Mikrodatorn som automatisk morseläsare. *Mikrodatorn* 3, 90-95.

Sigurd, B. 1984. Computer generation of spontaneous speech. COLING. Stanford.

Sinclair, J.M. 1985. Lexicographic evidence. In Ilson 1985, 81-92.

Sinclair, J.M. (ed). 1987a. *Looking up. An account of the COBUILD Project in lexical computing*. London: Collins.

Sinclair, J.M. 1987b. Grammar in the dictionary. In Sinclair 1987a, 104-115.

Sinclair, J.M. & M. Coulthard. 1975. *Towards an analysis of discourse: The English used by teachers and pupils*. Oxford: Oxford University Press.

Sinclair, J.M. & A. Renouf. 1988. A lexical syllabus for language learning. In Carter & McCarthy 1988, 140-160.

Stenström, A-B. 1984a. *Questions and responses in English conversation*. Lund Studies in English 68. Lund: Lund University Press.

Stenström, A-B. 1984b. Discourse tags. In Aarts & Meijs 1984, 65-81.

Stenström, A-B. 1986. What does *really* really do? Strategies in speech and writing. In Tottie & Bäcklund 1986, 149-163. [Also in Monaghan 1987, 65-79.]

Stenström, A-B. 1987. Carry-on signals in English conversation. In Meijs 1987, 87-119.

Stenström, A-B. 1988. Adverbial commas and prosodic segmentation. In Kytö, Ihalainen & Rissanen 1988, 221-236.

Stenström, A-B. 1989. Discourse signals: Towards a model of analysis. *Sprechen mit Partikeln*, ed H. Weydt, 561-574. Berlin: Walter de Gruyter.

Stubbs, M. 1983. *Discourse analysis: The sociolinguistic analysis of natural language*. Oxford: Blackwell.

Summers, D. 1988. English language teaching dictionaries: past, present, and future. *English Today* 14, 10-16.

Svartvik, J. 1980. *Well* in conversation. In Greenbaum, Leech & Svartvik 1980, 167-177.

Svartvik, J. 1987. Graphic English prosody. In Lindblad & Ljung 1987, 335-346.

Svartvik, J. 1988. English prepositions in applied linguistics. In J. Klegraf & D. Nehls, *Essays on the English language and applied linguistics*, 397-406. Heidelberg: Julius Groos.

Svartvik, J. & M. Eeg-Olofsson. 1982. Tagging the London-Lund Corpus of Spoken English. In Johansson 1982, 85-109.

Svartvik, J., M. Eeg-Olofsson, O. Forsheden, B. Oreström & C. Thavenius. 1982. *Survey of Spoken English: Report on research 1975-81*. Lund Studies in English 63. Lund: Lund University Press.

Svartvik, J. & R. Quirk (eds). 1980. *A corpus of English conversation.* Lund Studies in English 56. Lund: Lund University Press.

Svartvik, J. & A-B. Stenström. 1985. Words, words, words: The rest is silence? In Bäckman & Kjellmer 1985, 342-353.

Tannen, D. (ed). 1982. *Analyzing discourse: Text and talk.* Washington, D.C.: Georgetown University Press.

Thavenius, C. 1983. *Referential pronouns in English conversation.* Lund Studies in English 64. Lund: Lund University Press.

Tottie, G. 1989. What does *uh-(h)uh* mean? *Instead of flowers. Papers in honour of Mats Rydén on the occasion of his sixtieth birthday, August 27, 1989.* Acta Universitatis Umensis 90, eds B. Odenstedt & G Persson, 269-281. Stockholm: Almqvist & Wiksell.

Tottie, G. & I. Bäcklund (eds). 1986. *English in speech and writing: A symposium.* Studia Anglistica Upsaliensia 60. Stockholm: Almqvist & Wiksell.

Welin, C.W. 1979. *Studies in computational text comprehension.* Monograph 5, Institute of Linguistics, University of Stockholm.

# Index

LUND STUDIES IN ENGLISH
Founded by Eilert Ekwall
Editors: Sven Bäckman and Jan Svartvik

1. BERTIL WEMAN. 1933. Old English Semantic Analysis and Theory. With Special Reference to Verbs Denoting Locomotion. 187 pp.
2. HILDING BÄCK. 1934. The Synonyms for *child, boy, girl* in Old English. An Etymological-Semasiological Investigation. xvi + 273 pp.
3 GUSTAV FRANSSON. 1935. Middle English Surnames of Occupation. With an Excursus on Toponymical Surnames. 217 pp.
4. G. HOFSTRAND. 1936. *The Seege of Troye.* A Study in the Intertextual Relations of the Middle English Romance the *Seege or Batayle of Troye*. xv + 205 pp.
5. URBAN OHLANDER. 1936. Studies on Coordinate Expressions in Middle English. 213 pp.
6. VIKTOR ENGBLOM. 1938. On the Origin and Early Development of the Auxiliary *Do.* 169 pp.
7. IVAR DAHL. 1938. Substantival Inflexion in Early Old English. Vocalic Setems. xvi + 206 pp.
8. HILMER STRÖM. 1939. Old English Personal Names in Bede's History. An Etymological-Phonological Investigation. xliii + 180 pp.
9. UNO PHILIPSON. 1941. Political Slang 1759–1850. xvi + 314 pp.
10. ARTHUR H. KING. 1941. The Language of Satirized Characters in Poëtaster. A Socio-Stylistic Analysis 1579–1602. xxxiv + 258 pp.
11. MATTIAS T. LÖFVENBERG. 1942. Studies on Middle English Local Surnames. xlv + 225 pp.
12. JOHANNES HEDBERG. 1945. The Syncope of the Old English Present Endings. A Dialect Criterion. 310 pp.
13. ALARIK RYNELL. 1948. The Rivalry of Scandinavian and Native Synonyms in Middle English, especially *taken* and *nimen*. With an Excursus on *nema* and *taka* in Old Scandianvian. 431 pp.
14. HENNING HALLQVIST. 1948. Studies in Old English Fractured *ea.* 167 pp
15. GÖSTA FORSSTRÖM. 1948. The Verb *to be* in Middle English. A Survey of the Forms. 236 pp.
16. BERTIL WIDÉN. 1949. Studies on the Dorset Dialect. 179 pp.
17. CLAES SCHAAR. 1949. Critical Studies in the Cynewulf Group. 337 pp.
18 BERTIL SUNDBY. 1950. The Dialect and Provenance of the Middle English Poem *The Owl and the Nightingale.* A Linguistic Study. 218 pp.
19. BERTIL THURESSON. 1950. Middle English Occupational Terms. 285 pp.
20. KARL-GUNNAR LINDKVIST. 1950. Studies on the Local Sense of the Prepositions *in, at, on,* and *to,* in Modern English. 429 pp.
21. SVEN RUBIN. 1951. The Phonology of the Middle English Dialect of Sussex. 235 pp.
22. BERTIL SUNDBY. 1953. Christopher Cooper's *English Teacher* (1687). cxvi + 10* + 123 pp.
23. BJÖRN WALLNER. 1954. An Exposition of *Qui Habitat* and *Bonum Est* in English. lxxi + 122 pp.
24. RUDOLF MAGNUSSON. 1954. Studies in the Theory of the Parts of Speech. viii + 120 pp.
25. CLAES SCHAAR. 1954. Some Types of Narrative in Chaucer's Poetry. 293 pp.
26 BÖRJE HOLMBERG. 1956. James Douglas on English Pronunciation c. 1740. 354 pp.
27. EILERT EKWALL. 1959. Etymological Notes on English Place-Names. 108 pp.
28. CLAES SCHAAR. 1960. An Elizabethan Sonnet Problem. Shakespeare's Sonnets, Daniel's *Delia,* and their Literary Background. 190 pp.
29 ELIS FRIDNER. 1961. An English Fourteenth Century Apocalypse Version with a Prose Commentary. Edited from MS Harley 874 and Ten Other MSS. lviii + 290 pp.
30. The Published Writings of Eilert Ekwall. A Bibliography Compiled by Olof von Feilitzen. 1961. 52 pp.
31. ULF JACOBSSON. 1962. Phonological Dialect Constituents in the Vocabulary of Standard English. 335 pp.
32 CLAES SCHAAR. 1962. Elizabethan Sonnet Themes and the Dating of Shakespeare's Sonnets. 200 pp.33. EILERT EKWALL. 1963. Selected Papers. 172 pp.
34. ARNE ZETTERSTEN. 1965. Studies in the Dialect and Vocabulary of the *Ancrene Riwle.* 331 pp.
35. GILLIS KRISTENSSON. 1967. A Survey of Middle English Dialects 1290–1350. The Six Northern Counties and Lincolnshire. xxii + 299 pp.
36. OLOF ARNGART. 1968. The Middle English *Genesis* and *Exodus.* Re-edited from MS. C.C.C.C. 444 with introduction, notes and glossary. 277 pp.
37. ARNE ZETTERSTEN. 1969. The English of Tristan da Cunha. 180 pp.
38. ELLEN ALWALL. 1970. The Religious Trend in Secular Scottish School-Books 1858–1861 and 1873–1882. With a survey of the debate on education in Scotland in the middle and late 19th century. 177 pp.
39. CLAES SCHAAR. 1971. Marino and Crashaw. *Sospetto d'Herode.* A Commentary. 300 pp.
40 SVEN BÄCKMAN. 1971. This Singular Tale. A Study of *the Vicar of Wakefield* and its Literary Background 281 pp.
41. CHRISTER PÅHLSSON. 1972. The Northumbrian Burr. A Sociolinguistic Study. 309 pp.

42. KARL-GUSTAV EK. 1972. The Development of OE ȳ and ēo in South-Eastern Middle English. 133 pp.
43. BO SELTÉN. 1972. The Anglo-Saxon Heritage in Middle English Personal Names. East Anglia 1100–1399. 187 pp.
44 KERSTIN ASSARSSON-RIZZI. 1972. *Friar Bacon and Friar Bungay.* A Structural and Thematic Analysis of Robert Greene's Play. 164 pp.
45. ARNE ZETTERSTEN. 1974. A Critical Facsimile Edition of Thomas Batchelor. *An Orthoëpical Analysis of the English language* and *An Orthoëpical Analysis of the Dialect of Bedfordshire* (1809). Part I. 260 pp.
46. ERIK INGVAR THURIN. 1974. The Universal Autobiography of Ralph Waldo Emerson. xii + 288 pp.
47 HARRIET BJÖRK. 1974. The Language of Truth. Charlotte Brontë, the Woman Question, and the Novel. 152 pp.
48 ANDERS DALLBY. 1974. The Anatomy of Evil. A Study of John Webster's *The White Devil.* 236 pp.
49 GILLIS KRISTENSSON. 1974. John Mirk's *Instructions for Parish Priests.* Edited from MS Cotton Claudius A II and Six Other Manuscripts with Introduction, Notes and Glossary. 287 pp.
50 STIG JOHANSSON. 1975. Papers in Contrastive Linguistics and Language Testing. 179 pp.
51 BENGT ELLENBERGER. 1977. The Latin Element in the Vocabulary of the Earlier Makars Henryson and Dunbar. 163 pp.
52. MARIANNE THORMÄHLEN. 1978. *The Waste Land.* A Fragmentary Wholeness. 248 pp.
53. LARS HERMERÉN. 1978. On Modality in English. A Study of the Semantics of the Modals. 195 pp.
54. SVEN BÄCKMAN. 1979. Tradition Transformed. Studies in the Poetry of Wilfred Owen. 206 pp.
55. JAN JÖNSJÖ. 1979. Studies on Middle English Nicknames. I: Compounds. 227 pp.
56. JAN SVARTVIK & RANDOLPH QUIRK (eds). 1980. A Corpus of English Conversation. 893 pp.
57. LARS-HÅKAN SVENSSON. 1980. Silent Art. Rhetorical and Thematic Patterns in Samuel Daniel's *Delia.* 392 pp.
58 INGRID MÅRDH. 1980. Headlinese. On the Grammar of English Front Page Headlines. 200 pp.
59 STIG JOHANSSON. 1980. Plural Attributive Nouns in Present-Day English. x + 136 pp.
60. CLAES SCHAAR. 1982. The Full Voic'd Quire Below. Vertical Context Systems in *Paradise Lost.* 354 pp.
61 GUNILLA FLORBY. 1982. The Painful Passage to Virtue. A Study of George Chapman's *The Tragedy of Bussy D'Ambois* and *The Revenge of Bussy D'Ambois.* 266 pp.
62. BENGT ALTENBERG. 1982. The Genitive v. the of-Construction. A Study of Syntactic Variation in 17th Century English. 320 pp.
63. JAN SVARTVIK, MATS EEG-OLOFSSON, OSCAR FORSHEDEN, BENGT ORESTRÖM & CECILIA THAVENIUS. 1982. Survey of Spoken English. Report on Research 1975–81. 112 pp.
64. CECILIA THAVENIUS. 1983. Referential Pronouns in English Conversation. 194 pp.
65. NILS WRANDER. 1983. English Place-Names in the Dative Plural. 172 pp.
66. BENGT ORESTRÖM. 1983. Turn-Taking in English Conversation. 195 pp.
67. EVA JARRING CORONES. 1983. The Portrayal of Women in the Fiction of Henry Handel Richardson. 183 pp.
68 ANNA-BRITA STENSTRÖM. 1984. Questions and Responses in English Conversation. x + 296 pp.
69. KARIN HANSSON. 1984. The Warped Universe. A Study of Imagery and Structure in Seven Novels by Patrick White. 271 pp.
70 MARIANNE THORMÄHLEN. 1984. Eliot's Animals. 197 pp.
71. EVERT ANDERSSON. 1985. On Verb Complementation in Written English. 293 pp.
72. WIVECA SOTTO. 1985. The Rounded Rite. A Study of Wole Soyinka's Play *The Bacchae of Euripides.* 187 pp.
73. ULLA THAGG FISHER. 1985. The Sweet Sound of Concord. A Study of Swedish Learners' Concord Problems in English. xii + 212 pp.
74 MOIRA LINNARUD. 1986. Lexis in Composition. A Performance Analysis of Swedish Learners' Written English. x + 136 pp.
75 LARS WOLLIN & HANS LINDQUIST (eds). 1986. Translation Studies in Scandinavia. Proceedings from The Scandinavian Symposium on Translation Teory (SSOTT) II. 149 pp.
76 BENGT ALTENBERG. 1987. Prosodic Patterns in Spoken English. Studies in the Correlation between Prosody and Grammar for Text-to-Speech Conversion.
77. ÖRJAN SVENSSON. 1987. Saxon Place-Names in East Cornwall. xii + 192 pp.
78. JØRN CARLSEN – BENGT STREIJFFERT (eds). 1988. Canada and the Nordic Countries. 416 pp.
79 STIG CARLSSON. 1989. Studies on Middle English Local Bynames in East Anglia. 193 pp.
80. HANS LINDQUIST. 1989. English Adverbials in Translation. A Corpus Study of Swedish Renderings. 184 pp.
81. ERIK INGVAR THURIN. 1990. The Humanisation of Willa Cather Classicism in an American Classic. 406 pp.
82. JAN SVARTVIK (ED). 1990. The London-Lund Corpus of Spoken English Description and Research. 350 pp.

**Jan Svartvik (ed)**

# The London-Lund Corpus of Spoken English

## Description and Research

The appearance of this book marks the end of two projects in modern English linguistics, the London-Lund Corpus of Spoken English and Text Segmentation for Speech.

Part I is a description by Sidney Greenbaum and Jan Svartvik of the computerized London-Lund Corpus of Spoken English, which was completed in 1990. This product, offered to interested colleagues in all parts of the world, is the result of research, recording, analysis and compilation extending over many years and involving a great number of colleagues in the Survey of English Usage at University College London and in the Survey of Spoken English at Lund University. This part includes two appendices, one providing information about all the 100 texts of the corpus, the other listing publications that use material from the Survey of English Usage.

Part II reports on some of the research carried out within the Lund project Text Segmentation for Speech (TESS) by Bengt Altenberg, Mats Eeg-Olofsson, Anna-Brita Stenström, and Jan Svartvik.

Together with its companion volume in the same series, *A Corpus of English Conversation,* edited by Jan Svartvik and Randolph Quirk, this book will be a useful work of reference to those who use, or plan to use, the London-Lund Corpus. The new volume will also be of interest to those who are working on automatic text-to-speech conversion or are generally concerned with the analysis of computerized language texts.

Lund University Press
Box 141, S-221 00 Lund, Sweden

Art nr 20182
ISSN 0076-1451
ISBN 91-7966-126-2

Chartwell-Bratt Ltd
Old Orchard, Bickley Road,
Bromley, Kent BR1 2NE,
England

ISBN 0-86238-256-4