# ENGLISH IN SPEECH AND WRITING:
## A Project Report[1]

ANNA-BRITA STENSTRÖM

*University of Lund*

## *1. Introduction*

The research project The Survey of Spoken English ended in the spring 1981, and a new project, English in Speech and Writing, with the Swedish acronym ETOS, was started in 1981 as a follow-up of the previous one. The director of the new project, which receives financial support from the Swedish Research Council for the Humanities and Social Sciences, is Gunnel Tottie.

The aim of ETOS is to achieve an explicit description of important differences between spoken and written English, ie. a description which will contrast the use and non-use of various linguistic features in the two media. In the descriptions of spoken language presented by eg. Crystal and Davy (1975) and Brown (1977) the contrast with written language is present, but more often than not it is implicit, ie. differences between speech and writing are not spelled out in detail. It is taken for granted that eg. expressions like *you know* or *well* are speech-specific, and their uses are described in some detail, but the authors do not discuss why these and other expressions are speech-specific, or what written language uses instead, if anything.

ETOS is not the first research project contrasting spoken and written language. Earlier projects at Lund University have dealt with Swedish language material, and there are several ongoing projects in the United States, eg. at Berkeley University, California, where Wallace Chafe has studied 'maximal differentiated styles: informal spoken language and formal written language' (Chafe 1982).

[1] I am indebted to the following persons who have not only commented on the report but also contributed with material and provided results from their own research: B. Altenberg, L. Hermerén, J. Svartvik and G. Tottie.

Apart from these projects, comparatively few earlier studies have been devoted to the differences between speech and writing. For lack of adequate data, especially from spontaneous spoken discourse, most of these have been limited in aim and scope and the results have been fairly trivial (for references see Tannen 1982).

## 2. *Research goals*

What characterizes ETOS is that we take a functional approach in the full sense of the word, addressing problems concerning the grammatical as well as pragmatic functions of linguistic items. The research goals of ETOS can be summarized under the following headings:

- *From form to function*
  What function(s) — grammatical or pragmatic, does a particular item (or class of items) fulfil in speech and writing? For instance, how are negative expressions used? How are modal verbs or logical connecters used?
- *From function to form*
  Which linguistic items can be used to fulfil a particular function? For instance, what do adverbials look like in speech and writing? Do they consist mostly of adverbs (as *often, there, why*), of prepositional phrases (as *on Monday, in the bin, for what reason*) of or adverbial clauses (as *when I saw him, where you were sitting, because I was angry*)? How are modal meanings expressed? Are they expressed mostly by modal verbs (as *he can't pay his debts*) or by other means, (as in *he has no means of paying his debts*)?
- *Quantification*
  What differences are there in the functions required by speech and writing? It seems obvious that there must be more questions in speech, but what about negation, modality, adverbials? Is there more or less of any of these categories, and are the types of negation, modality or adverbial expressions the same or different in the two varieties? It would seem natural to expect more expressions of obligation and permission in speech than in writing, but what are the proportions? And what about the other modalities?
- *Information structure*
  How are speech and writing organized, respectively? We know that well-organized sentences of the standard grammar-book type are hard to find in spoken language. But what exactly is it that characterizes spoken language then? How is information conveyed, and by means of what structures?

- *Explanation*
  Ultimately, we want to know the reasons for the differences we find. Are they due to the differences in communicative situations between speech and writing, or to psychological factors, either on the production or the perception side? These are complicated matters where we can only hope at present to come up with plausible hypotheses and educated guesses.

## 3. *Spoken and written material*

It is clear that there are many variants of spoken and written communication, variants which may be regarded as points on a scale ranging from 'most typical' to 'least typical' of each medium. We have chosen to study what we deem to be the most typical variants of each medium, viz spontaneous conversation and non-fictional informative prose. It seemed to us that these variants would provide the most fruitful contrasts. They are maximally contrasted not just through medium but because of the situations in which they are used, as well as the purposes they serve. Conversation is used in human interaction with at least two participants, and its purpose is not normally restricted to the acquisition or imparting of information. We also indulge in conversation to fulfil our need for social interaction with other human beings (phatic communion) conveying at the same time our attitudes and emotions by means of gestures, voice quality, etc. Usually, too, we can rely to a large extent on the situational context to provide clues to what we mean, and we therefore often need to be less explicit than when we write. Moreover, when we converse, we are normally pressed for time, in the sense that we plan and produce our linguistic output simultaneously.

The situation is vastly different when we communicate in writing, especially when we produce informative prose. Our purpose is precisely to inform. We usually have time to plan our message carefully before committing it to the written word. The recipient of the message is not normally present, and we cannot therefore rely on the situational context to provide him with clues concerning meaning. We are thus forced to be more explicit, and to express our message in the clearest possible terms, as we cannot check how it is received.

Choosing conversation and informative prose to represent the spoken and written media also had practical advantages. Large collections of linguistic material are necessary to carry out the kinds of research that we wish to undertake, and we are fortunate enough to have access to such collections of material stored on computer tape. The corpora we work with are the London-Lund Corpus of English Conversation, abbreviated LLC (published in part in Svartvik and Quirk 1980), and the Lancaster-Oslo/Bergen Corpus, abbreviated LOB. Both of these corpora contain exclusively British English of comparatively recent date: LLC is based on recordings mainly from the sixties, and LOB contains printed material from 1961.

### *3.1 The London-Lund Corpus of Spoken English*

LLC is a corpus of educated spoken British English. It is part of the large (spoken and written) material collected — chiefly in the 1960s — at the Survey of English Usage under the direction of Randolph Quirk, University College London (App 1). In 1975, the spoken material, which had been analysed prosodically and transcribed on paper slips in London, was put at the disposal of the Survey of Spoken English under Jan Svartvik at Lund University. The material has been transferred to computer tape and is now available for further analysis in machine-readable and printed form. The corpus comprises 87 texts of about 5.000 words each, or almost half a million words in all, and represents a variety of speech situations (conversation, radio interviews, public speeches, etc (see App 2)).

LLC is available for research in three versions:

1. magnetic tape for computer processing
2. printed version of running text: subgroup A (in Svartvik and Quirk (1980); see App 3)
3. KWIC concordance on computer tape: subgroups A—H

The printed version of the running text consists of surreptitiously recorded conversation (subgroup A: 34 texts comprising about 170.000 words; App 2—3). The concordance is also available at the Survey in a printout copy. There are also printouts of alphabetical and rank-ordered frequency lists.

For details on the corpus, see Svartvik and Quirk (1980) and Svartvik et al. (1982).

### *3.2 The Lancaster-Oslo/Bergen Corpus of British English*

LOB is a British English equivalent of the Brown Corpus (BC), which is a collection of American English produced at Brown University, Providence, Rhode Island, under the direction of Nelson Francis. BC is exclusively drawn from printed sources published in 1961 and comprises 500 different text samples of about 2.000 words each, representing 15 different categories or genres (press, reportage, religion, science, fiction, humour, etc.). In all, the corpus contains approximately one million running words (see App 4). LOB was initiated by Geoffrey Leech at Lancaster University, England, and completed and prepared for computer analysis by Stig Johansson, Oslo University, and the Norwegian Computing Centre for the Humanities at

Bergen. It was designed to match BC and is consequently, as far as possible, comparable to its American predecessor as regards size, year of publication, and sampling principles (see App 4). LOB is available in the following versions:

1. magnetic tape for computer processing (LOB TAPE)
    a) printout (LOB TEXT)
2. running text
    b) microfiche (LOB FICHE)
3. KWIC concordance: microfiche (LOB KWIC FICHE)
4. word frequency lists: printed in Hofland & Johansson 1982 (LOB REVERSE)

Detailed information on the corpus is given in a manual (Johansson et al. 1978) and in Hofland and Johansson (1982). The latter contains a statistical analysis of the vocabulary in LOB — comparable to that of BC in Kučera and Francis (1967) — including alphabetical and rank-ordered frequency lists, word frequencies in different text categories, and (on micro-fiche) a reverse-alphabetical word list. Hofland and Johansson also contains a comparison of word frequencies in LOB and BC and is, in fact, a valuable source of information for comparative studies of British and American vocabulary

### *3.3 Mini and Midi corpora*

For the purposes of ETOS, the two standard corpora — LLC and LOB — are generally too large and unwieldy to be investigated in their entirety. For this reason two smaller 'project corpora' were selected from the larger ones. They are referred to as the Mini and Midi corpora.

As the names indicate, the Midi corpus is larger (2 × 100.000 words) than (and includes) the Mini corpus (2 × 10.000 words), but both types are otherwise composed according to the same principles, viz to represent an equal amount of conversational spoken English (from LLC) and informative written English (from LOB).

The Mini corpus is intended for pilot investigations and other limited studies that do not require a large material, whereas the Midi corpus is better suited for more extensive studies.

### *3.4 Comparability*

Something should be said about the comparability of the spoken and written material, of LLC and LOB. The two corpora were not originally designed for comparative work, but we nevertheless feel justified in using them for this purpose. The speakers taking part in the conversations of LLC are for the most part academics with a background in the humanities, and the non-fictional texts of LOB which are used for the purposes of the project are precisely the kind of texts that LLC speakers might be expected to produce: ie. they are not examples of highly specialized technical or scientific writing but of a journalistic or essayistic type.

## 4. *Current work within the project*

Within ETOS, research is currently being carried out along the following lines:

Gunnel Tottie is working on problems of negation in English, especially
the variation between the types exemplified by *He saw nobody* and *He did not see anybody;*
the pragmatics of negation, especially factors conditioning the frequency and occurrence of different functions of negative expressions, eg whether they occur as responses to questions, as denials of previous statements, as rejections of offers, etc.

She is also working on
the use of adverbials in spoken and written language, their frequency of occurrence, different functions (eg as adverbials of manner, time, etc.), and their different types of realization, as adverbs, prepositional phrases, or clauses.

Lars Hermerén is working on the expressions of modality in speech and writing, especially two problems:
the extent to which modality is expressed by means of modal verbs and the extent to which it is expressed by other means the frequencies of different types of modalities and their expressions (eg Certainty and Belief, expressed for instance by *must* and *think,* and Necessity and Possibility, which may be expressed by *must* and *perhaps,* respectively)

Jan Svartvik is working on the relation between grammar and prosody, or intonation structure, and is studying especially the following phenomena:
word-class distribution;
the structure of grammatical phrases, eg the complexity of noun phrases and verb phrases;
the structure and content of tone units;
planning spans and hesitation phenomena;

Together with Mats Eeg-Olofsson he is also working on tagging, which is of considerable importance for the future use of corpora such as LLC. By 'tagging' is meant the assignment of lexical or grammatical categories to items in the corpus (eg. noun and verbs, subjects and complements) and the labelling of the respective items with the appropriate tag. If a corpus is properly tagged, it will be possible to extract information that is not readily available otherwise. A great deal of tagging has already been carried out within the project Survey of Spoken English, and the system of tagging is described in Svartvik et al. (1982).

Within the last year, the levels of tagging have been extended and now comprise

1. Word-class (main verb, preposition, etc.), eg:
   *'ll* ⟨VM+8⟩
   *be* ⟨VB+0⟩
   *seeing* ⟨VA+G⟩
2. Grammatical phrase (verb phrase with the verb in the present tense, plural noun phrase, etc.), eg:
   *'ll be seeing* ⟨VPH: modal progressive⟩
3. Clause element (subject, complement, etc.), eg:
   *I'll be seeing her* ⟨S V O⟩
4. Discourse element (softener, greeting, etc.), eg:
   *I'll be seeing her you know*
   x x x SOFT

This four-level tagging system is currently applied to the text, tone unit by tone unit, but it is envisaged that it will eventually be extended to include also adjacent tone units or longer sequences of tone units. For each level, Jan Svartvik has written a set of algorithms which have been translated into the programming language Simula by Mats Eeg-Olofsson.

For the word-class level, Mats Eeg-Olofsson has also worked out a method to achieve automatic word-class tagging based on frequencies of tags and tag combinations in tone units and on the identification of items by graphemic patterns, ie. the sequences of different letters which are characteristic of different word-classes. Thus, for instance, *—ion* and *—er* are typical endings of nouns.

For further information on the tagging system, see Svartvik and Eeg-Olofsson (1980), Svartvik et al. (1982) and Svartvik (1982).

Bengt Altenberg is working on a comparative study of logical connecters (*yet, although, so, therefore,* etc.). Using a sample of surreptitiously recorded conversation and a sample of informative prose, each amounting to c 100.000 words, he has so far examined three aspects of causal connection:

— the choice of connecter in the two samples
— the syntactic type of linkage involved:
parataxis (clauses at the same level linked by an adverbial connecter);
hypotaxis (clauses at different levels linked by a subordinater);
clause integration (the connective expression is fully integrated as subject or complement in the clause structure; *the reason is, that's why,* etc.)
— the order of the related propositions (cause — result: CR order; result — cause: RC order)

In all, the material was found to contain 1.173 connective expressions, representing 66 different 'realization types'. Of these types, 58 were employed, in the written sample and 38 in the spoken sample. In terms of tokens, however causal connecters were almost twice as frequent in the spoken as in the written material. In other words, although overt expressions of causal relations were much more frequent in the spoken discourse, they were more stereotyped. This is highlighted by the fact that the subordinater *because* and the conjunct *so* together accounted for 79% of the tokens in the spoken sample, but only 23% in the written sample, which instead made greater use of *for, therefore, since* and *thus*.

The spoken and written samples were also found to differ slightly as regards the sequence of the cause-result relation, speech preferring RC order and writing CR order. A possible reason for this may be that the CR sequence, although it reflects a 'real world' ordering of causal events, requires a greater amount of planning. Since spontaneous conversation is typically unplanned, a postposed cause or reason (RC order) may be easier to process in impromptu speech.

Both the spoken and the written samples were found to prefer hypotactic constructions to paratactic ones, which were in turn much more common than clause-integrated expressions. The major difference between the two media was that, while the spoken sample showed a somewhat greater preference for the first two types, clause integration was on the whole more common in the written sample.

## 5. *Other studies based on the project corpora*

The following articles and term papers are based on written and spoken material from the LLC and LOB corpora.

In an article entitled "The missing link? Or, why is there twice as much negation in spoken English as in written English?" Gunnel Tottie has tried to account for the higher number of negations in speech. Studying the incidence of negation in spoken and written English in two samples of 50.000 words each, she found that negation occurred twice as frequently in speech. On the basis of a pragmatic theory formulated in Tottie (1982) she suggested that one plausible reason for the difference is the existence in conversation of two kinds of negation, rejections and explicit denials, which do not exist in written language, where only implicit negation occurs.

She tested the hypotheses by examining a subset of the 50.000-word spoken sample and found that explicit denials did not account for more than 16% of the total number of negatives in the three texts examined. The use of negatives in other speech-specific categories, such as direct questions, feedback signals, and imperatives, accounted for another 17%, etc. When

all had been accounted for there was still a gap of 16% that could not be explained.

Therefore Gunnel Tottie decided to pursue another line of investigation, namely the cooccurrence of negation with modal and mental verbs. It had been observed in other studies (Svensson 1981) that negative expressions tended to cooccur with both modal and mental verbs. Moreover, Chafe (1982) had found that spoken language contained a higher frequency of references to speakers' mental processes than written language.

In their term papers, two third-term students of English (A. Bengtsson and M. Bertilsson) showed that spoken and written samples had very similar proportions of modals and that modals occurred more frequently in negative than in non-negative sentences in both samples. These findings were consequently not very helpful. The findings with respect to mental verbs were much more favourable and showed that mental verbs occurred more frequently in spoken than in written language and also that they manifested a high tendency to collocate with negation. So, part at least of the missing link was found in the shape of collocations of mental verbs with negation.

Starting from Crystal's (1980) claim that adverbials are more or less necessary in conversational clause structure as compared with written sentences, where they have traditionally been considered optional, Ulla Hedling wrote a term paper on "The frequency of adverbials in written and spoken English". She limited her study to adverbials answering questions introduced by *When, How long, How many times, Where* and *How* in a sample of 10.000 words, half of which consisted of informal speech from the LLC corpus and the other half of equivalent texts from the LOB corpus.

Except for the general tendency in speech to avoid complexity, here manifested in a preference for one-word adverbials to noun phrases and prepositional clauses, she found that:

- there were almost twice as many expressions of TIME in spoken as written language. *When*-responses dominated, mostly realized by *now, then,* and *just,* or noun phrases, such as *this year* and *last term*
- PLACE adverbials occurred twice as frequently in written language and were mainly expressed by prepositional phrases, eg. *at Halidon Hill*
- MANNER adverbials were three times as frequent in written language, mostly expressed by adverbs in writing but by prepositional phrases in speech. The comparatively high figure for manner adverbials in writing may be explained by the greater need for clarity and fear of ambiguity in the written medium, where no interaction takes place
- DEGREE adverbials, especially premodifying intensifiers, were much more common in speech. The most common intensifier was *very*, followed by *quite*

Ulla Hedling concludes by saying that the somewhat higher percentage of adverbials in the spoken material seems to support Crystal's theory to some extent, although her study is based on a very small sample indeed.

Another third-term student, Mats Johansson, studied the complexity of spoken and written language as it is manifested in the use of subclauses in two samples of 5.000 words each, one spoken and one written, from LLC and LOB respectively. In order to get his samples as comparable as possible he excluded clauses consisting of *you know* and *you see*, acting as conversational fillers, Q-tags, greetings, and broken-off utterances which only occurred in the spoken material.

Judging by the two samples he found two major differences between speech and writing with respect to the use of subclauses. One was that speech seems to use *that*-clauses with object function to a greater extent that writing and the other that writing seems to contain considerably more relative clauses than speech.

He tries to explain these findings. In the first case, it may be possible, he says, that the abstract reference to *that*-clauses is carried out differently in writing, eg. by a more frequent use of abstract noun phrases. In the second case, the difference may be due to a wish to avoid repetition in writing. He develops this further and suggests two possible explanations for the fact that relative clauses were almost three times as common in written English:

1. there might be more noun phrases in writing and therefore more opportunities for postmodification
2. the number of noun phrases may be the same in both samples but more complex in writing. If this is true and if it can be assumed that nouns are modified equally often in both samples, there may be a difference in modifying technique. Spoken language tends to keep main clauses short, which may be taken to indicate that nouns in speech are modified in a separate main clause instead of by a subclause, as illustrated in (slants indicate tone unit boundaries):
   a Stoke student has made a copy of the painting which/the painting's in Madrid/ I think/ it's not in London/ which seems to reflect that the speaker is choosing between postmodification with a relative clause and a less complex construction with two main clauses.

Mats Johansson's conclusion is that the tendency towards compactness and avoidance of repetition in written language does not seem to be matched in spoken language.

Drama is said to reflect real life and real characters. On this basis Zigmar Fritzon decided to investigate to what extent features that characterize casual everyday conversation were used to create the effect of real life conversation in drama dialogue. He studied such features as 1) softening connectives: *you know, you see, I mean, mind you, sort of* and *kind of,* 2) Q-tags, and 3) minor sentences in four plays, two by Pinter and two by Ayckbourn, which he compared with two LLC texts. The plays were carefully picked out so that the language in the two samples corresponded, ie. could be described as educated speech.

He found that softening connectives were more than three times as frequent in the spoken sample, that Q-tags occurred about twice as often in speech as in writing but that minor sentences were more common in drama dialogue.

Among the softening connectives *you know* was considerably more common than the rest but occurred much less frequently when the speaker was talking about a subject in which he was well versed and had no need to pause for thought. In one of the texts a speaker employed numerous *you know* when talking about a delicate matter but stopped using the device when he was back on neutral ground.

*Kind of*, which is more typical of American than British English, was rare. *Sort(kind) of* occurred only twice in the drama dialogue but 31 times in the spoken texts. The reason for this may be that this device creates an impression of vagueness, undesirable in a drama dialogue which is supposed to give a clear picture both of the characters and the plot.

Referring to Crystal and Davy (1975), Zigmar Fritzon states that both *I mean* and *sort/kind of* indicate that the speaker assesses the conversation as informal. But whereas *I mean* expresses the speaker's attitude both to the listener and to what he is saying, *sort/kind of* rather expresses his attitude to what he is saying.

One of the functions of Q-tags, he says, is to keep the conversation going by ensuring active participation of the listener, but Q-tags and softening connectives often have the same function and are therefore often interchangeable.

As to minor sentences, the option to use *m* and *yeah* to express one's opinion is available in speech but is not used in drama dialogue, where adverbials, such as *quite, really* and *super* are used instead. Common to both samples was the omission of the subject.

Summing up, Zigmar Fritzon found that the realism of the plays could be detected in the language only to some extent. He found that if the drama dialogue had contained as many softening connectives as the spoken texts, this would have created an impression of disjointedness and non-fluency. The degree of formality in a drama is generally established early on, and consequently there is no need for softening connectives for that reason.

With respect to Q-tags, the dramatists succeeded in matching spoken language. Apparently, Q-tags are more accepted as a feature of speech than softening connectives. One proof of this is that they are much more extensively covered in grammars.

Finally, it is doubtful, he says, whether one should compare drama and natural conversation in terms of minor sentences; after all, written language is divided into sentences separated by punctuation, whereas speech is divided into tone units separated by intonation.

## 6. *Concluding remark*

The financial support of the ETOS project will come to an end in June, 1984, but work on speech and writing will be going on.

App 1. Composition of material in the Survey of English Usage (from Svartvik & Quirk 1980).

(I) Material with origin in writing (100 texts)

| (A) Printed (46) | | | (B) Non-printed (36) | | |
|---|---|---|---|---|---|
| Learned arts | | 6 | Continuous writing | imaginative | 5 |
| Learned sciences | | 7 | | informative | 6 |
| Instructional | | 6 | | | |
| Press | general news | 4 | Letters: social | intimate | 6 |
| | specific reporting | 4 | | equal | 3 |
| | | | | distant | 4 |
| Administrative & official material | | 4 | | | |
| Legal and statutory material | | 3 | Letters: non-social | equal | 4 |
| Persuasive writing | | 5 | | distant | 4 |
| Prose fiction | | 7 | Personal journals (diaries) | | 4 |

(C) As Spoken (18)

| | | |
|---|---|---|
| Drama | | 4 |
| Formal scripted oration | | 3 |
| Broadcast news | | 3 |
| Talks | informative | 4 |
| | imaginative | 2 |
| Stories | | 2 |

(II) Material with origin in speech (100 texts)

(A) Monologue (24)

| | | | |
|---|---|---|---|
| Prepared (but unscripted) oration | | | 6 |
| Spontaneous | oration | | 10 |
| | commentary | sport | 4 |
| | | non-sport | 4 |

(B) Dialogue (76)

| | | | |
|---|---|---|---|
| Conver-sation | Surrepti-tious | intimate | 24 |
| | | distant | 10 |
| | Non-surrepti-tious | intimate | 20 |
| | | distant | 6 |
| | Tele-phone | intimate | 11 |
| | | distant | 6 |

App 2. Text classification of the London-Lund Corpus of Spoken English (LLC) (from Svartvik ar el. 1982).

| TEXT GROUPS | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| ABBRE-VIATION | BROAD DESCRIPTION | Subgroups | Dialogue | Face-to-face | Private | Surrepetitious | Radio | TEXT LABELS | SUB-GROUP TOTAL | TEXT GROUP TOTAL |
| CON | Face-to-face conversation | A | + | + | + | + | − | S. 1. 1−14, S. 2. 1−14, S. 3. 1−6 | 34 | |
| | | B | + | + | + | − | − | S. 4. 1−7, S. 5. 8−11, S. 6. 2[1]) | 12 | 46 |
| TEL | Telephone conversation | C | + | − | + | + | − | S. 7. 1−3, S. 8. 1−4, S. 9. 1−3[2]) | 10 | 10 |
| DIS | Discussion, interview, debate | D | + | + | − | − | + | S. 5. 1−7, S. 6. 1, S. 6. 3−5[3])<br>S. 10. 8[4]) | 12 | 12 |
| PUB | Public, unprepared commentary, demonstration, oration | E | + | + | − | − | − | S. 11. 1, S. 11. 4−5 | 3 | |
| | | F | − | + | − | − | + | S. 6. 6[5]), S. 11. 3 | 2 | 12 |
| | | G | − | − | − | − | + | S. 10. 1−7[6]) | 7 | |
| PRE | Public, prepared oration | H | − | + | − | − | − | S. 11. 2, S. 12. 1−6[7]) | 7 | 7 |
| | | | Number of "+" texts: | | | | | | | |
| | | | 71 | 70 | 56 | 44 | 21 | | 87 | 87 |

App 3. LLC TEXT A: Extract from the running text of LLC (subgroup A) (from Svartvik & Quirk 1980).

S.1.13
AS | SÌSTANT — [42] | couldn't ΔPÒSSIBLY■ [43] | have a WÓMAN [RÉALLY]■
[44] |being a DÉNTIST■ [45] it's [n] it's | not all that △ÈASY you SÉE■ — . [46] and
| I said [n] | WÈLL■ [47] [ə] I said | well I △know that the EX'ÂMS are DÍFFERENT■
— [48] and that [ə] | if you ΔWÊNT to A'merica■ [49] you'd | have to ΔTĂKE a 'dental
ex'am■ [50] be|fore you could ΔPRÀCTISE■ [51] but I said | Pam had '△several of
△HÈR con'temporaries■ — [52] who |went ÒVER■ [53] to | either ΔCÀNADA or the
STÁTES■ [54] in |order to'make some△quick'money to △come'back and △set up'on their
△ŎWN■. [55] and | they had | didn't have any'△difficulty in'△PĂSSING «the» . the
ex'am■. [56] so |Eileen was' VÊRY im'pressed■ [57] she | said 'well ΔYÈS■ [58] «she said»
|that's ΔVÈRY interesting■. [59] [?] . |I ΔKNÒW■ [60] that |when . ΔPĂM had'qualified■
[61] there were | very △FÈW women'dentists in [A|MÈRICA■]■ . [62] and I | said well
Δquite FRĂNKLY■ [63] a|bout a ΔTHÌRD of'all the △[|DÈNTAL 'students■]■ [64] in
| ÊNGLAND * NÓW■* [65] | are ΔWÒMEN■ —
C [66] *|[m̀]■*
a [67] well it surprises me that Eileen should be surprised I can imagine Leslie being
surprised but America — she must know that — lots of dentists who are women —
B [68] «remember» 'why SHÔULD'she■ [69] cos she | hasn't ΔLÌVED in'England NÓW■
[70] for — | thirty — . * 'odd YÉARS■*
a [71] *no but she lived* here for twenty years «4 sylls»
B [72] '| oh NÒ■. [73] there were | very FÈW■ [74] «you know [əm]» |VÊRY 'few'women■
[75] [əm]. |you SÉE■ [76] |women △DĔNTISTS■ [77] in the |days when you △had to
△PĂY■ [78] would |never have'made a ΔLÌVING■ [79] because |NÓBODY would have
GÒNE to a'woman'dentist■ — [80] it |was be'cause of the ΔSHÒRTAGE■. [81] '|why
'women are △so ac'cepted NÓW■ [82] |[?]is■ [83] that be|cause of the ΔSHŎRTAGE
of'dentists [in |ÉNGLAND■]■ [84] '|all the ΔSCHŎOL'dental'officers■ [85] | after the
WĂR■ [86] | were — be|cause they were 'badly PÁID■ [87] | were the ΔWÒMEN■
[88] *|when they . |when they
a [89] *I'm quite I'm sure*
>B [88] QUÁLIFIED■* [90] |they be'came — ΔSCHÒOL'dental ▷officers■ [91]* and |so «there
was»*
a [92] *I'm sure there* was a school dental officer «who was» a woman when I was a child
B [93] |WÈLL■ [94] |we never ΔWÈNT to school'dental'officers■ [95] so I |wouldn't have
ΔKNÒWN■ [96]*but they |wouldn't have*
a [97] *well neither did I * but you just got done «I mean»

App 4. The basic composition of BC and LOB (from Hofland & Johansson 1982)
Text categories A—J =informative prose
H—R=imaginative prose

| Text categories | | Number of texts in each category | |
|---|---|---|---|
| | | American corpus | British corpus |
| A | Press: reportage | 44 | 44 |
| B | Press: editorial | 27 | 27 |
| C | Press: reviews | 17 | 17 |
| D | Religion | 17 | 17 |
| E | Skills, trades, and hobbies | 36 | 38 |
| F | Popular lore | 48 | 44 |
| G | Belles lettres, biography, essays | 75 | 77 |
| H | Miscellaneous (government documents, foundation reports, industry reports, college catalogue, industry house organ) | 30 | 30 |
| J | Learned and scientific writings | 80 | 80 |
| K | General fiction | 29 | 29 |
| L | Mystery and detective fiction | 24 | 24 |
| M | Science fiction | 6 | 6 |
| N | Adventure and western fiction | 29 | 29 |
| P | Romance and love story | 29 | 29 |
| R | Humour | 9 | 9 |
| Total | | 500 | 500 |

## REFERENCES

Altenberg, B. "Casual connection in spoken and written English". *Studia Linguistica* 38/1. 20—69.

Bengtsson, A. 1982. "Modal auxiliaries, mental verbs and subject types in negative clauses. A study of written English". Unpublished term paper.

Bertilsson, M. 1982. "Modal auxiliaries, mental verbs and subject types in a sample of of spoken English". Unpublished term paper.

Brown, G. 1977. *Listening to spoken English.* London: Longman.

Chafe, W. 1982. "Integration and involvement in speaking, writing, and oral literature". In Tannen, D. (ed.). 1982. 35—53.

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

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

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

Fritzon, Z. 1982. "A comparative study of spoken English and dialogue in drama". Unpublished term paper.

Hedling, U. 1982. "Adverbials in speech and writing". Unpublished term paper.

Hofland, K. and Johansson, S. 1982. *Word frequencies in British and American English.* Bergen: Norwegian Computing Centre for the Humanities .

Jacobson, S. (ed.). 1983. *Proceedings from the Second Scandinavian Symposium* on

*Syntactic Variation at Stockholm University*. Stockholm: Stockholm Studies in English 57.

Johansson, M. 1982. "Non-finite clauses in speech and writing". Unpublished term paper.

Johansson, S. (ed.). 1980. *Computer corpora in English language research.* Report from the Norwegian Computing Center for the Humanities. Bergen.

Johansson, S., Leech, G. and Goodluck, H. 1978. *Manual of information to accompany the Lancaster-Oslo/Bergen corpus of British English, for use with digital computers.* Oslo: Department of English, University of Oslo.

Kučera, H. and Francis, W. N. 1967. *Computational analysis of present-day American English.* Providence, R. I.: Brown University Press.

Svartvik, J. 1982. "The segmentation of impromptu speech". In Enkvist, N. E. (ed.). 1982. 131–45.

Svartvik, J. and Quirk, R. 1980. *A corpus of English conversation.* Lund: Gleerup.

Svartvik, J. and Eeg-Olofsson, M. 1980. "Tagging the London-Lund corpus of spoken English". In Johansson, S. (ed.). 1980. 85–109.

Svartvik, J., Eeg-Olofsson, M., Forsheden, O., Oreström, B. and Thavenius, C. 1982. *Survey of spoken English. Report on research 1975–81.* Lund Studies in English 63. Lund: Gleerup.

Svensson, J. 1981. "Etermediernas nyhetsspråk 2. Studier över innehåll och informationsstruktur". *Lundastudier i nordisk språkvetenskap ser C, No. 11.* Lund: Walter Ekstrand.

Tannen, D. (ed.). 1982. *Spoken and written language. Advances in discourse processes.* Vol. IX. Norwood, N. J.: Ablex.

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"?. In Jacobson, S. (ed.). 1983. 67–84.