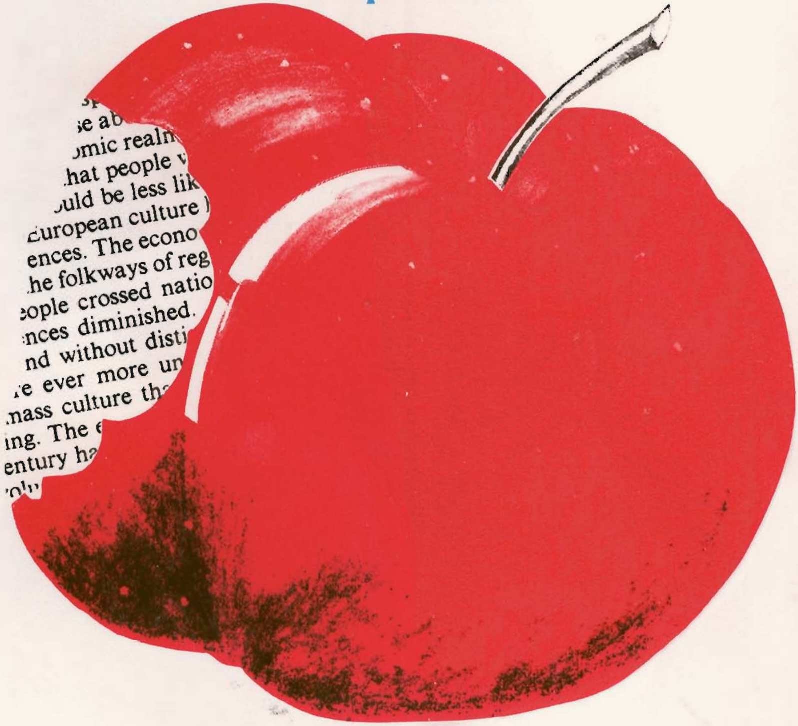
ICAME Journal

International Computer Archive of Modern English

No. 15

April 1991



NORWEGIAN COMPUTING CENTRE
FOR THE HUMANITIES

Contents

Articles:

Reviews:

Shorter notices:

The *ICAME Journal* is the continuation of *ICAME News.*
Editor: Stig Johansson, University of Oslo

Sequences of Temporal and Spatial Adverbials in Spoken English: Some Pragmatic Considerations

Hilde Hasselgård
University of Oslo

Introduction

In this paper I am examining two categories of spoken English; Conversation and Commentary. One aim is to clarify how these text types differ from one another in various respects, and to what extent text type has any influence on adverbial placement and the organization of adverbial sequences. Section 1 gives a very brief introduction to the classification scheme used in my investigation. After a discussion of the text types examined, I am trying out the relevance of this kind of text typological analysis on the language material. For this purpose I have chosen to examine more closely the Commentary category.

1. The classification of adverbial sequences

An adverbial sequence denotes two or more adverbials which belong to the same clause. I am examining adverbial sequences containing two or more spatial and/or temporal adverbials.

I distinguish two main types of adverbial sequences. A sequence of adverbials which occur in the same position is termed a *cluster*. Spatial/temporal adverbials in a cluster are normally adjacent, but may be separated by some other optional constituent, e.g. an adverbial other than time or space.

(1) Celia came *to London a little over a year ago.*

(1a) Celia came *to London* on her own *a little over a year ago.*

The other type of sequence is called a *combination.* Combinations are sequences where two (or more) adverbial slots have been filled.

(2) *On the 6th* winds gusted to 90 mph *at Mumbles in South Wales.*

The two types differ in frequency. Clusters are the more frequent type in all the text categories examined so far. This suggests that when adverbials of time/space occur in the same clause, they tend to be attracted to each other. Clusters can thus tentatively be viewed as more basic than combinations.

In clusters in End position (where most of the clusters occur) there is a marked tendency for the adverbials to follow a certain order, the main principle being that spatial adverbials precede temporal ones. But in combinations no fixed order has been found. There is some evidence that combinations are used when the speaker/writer wants to use a sequence of adverbials that cannot be realised as a cluster, or which is very rarely realised as such.

In combinations the time + space order is the more frequent one, whereas clusters show the opposite tendency. This difference in structure is partly due to the fact that while spatial adverbials tend to be placed in End position, temporal adverbials are more apt to occupy other positions as well.

2. The material for the investigation

The source material for the present paper has been taken from the London-Lund Corpus of Spoken English, which is a sub-section of the *Survey of English Usage* material collected at University College London. The texts in the corpus are sorted according to text types. I have selected texts from the categories "conversation" and "commentary". The former category consists of surreptitiously recorded conversations. Most of the speakers are academics, or have some connection with University College, where most of the recordings took place. The commentary category can be further subdivided into (a) sports commentaries and (b) radio/TV commentaries on "non-sport" public events. The conversation material examined here consists of 19 texts, the commentary material of 8. Each text totals 5000 words.

For my purpose I have extracted the sentences (or what should be roughly equivalent to sentences in speech) which contain a sequence of temporal and/or spatial adverbials. Yet I propose to examine those sentences in a wider context, because I find that the factors governing the placement of adverbials and the organisation of adverbial sequences are often to be found in the surrounding text. This is also the reason why I came to suspect that there might be differences between text types as to the organisation of adverbial sequences.

3. The text types examined

Biber (1986) emphasises the importance of distinguishing between different types of spoken as well as written language. He claims that contradictory results of comparisons between speech and writing can be ascribed to the fact that speech and writing have been "considered as coherent wholes, and findings [have] typically [been] presented as general linguistic characteristics of a single dimension in which the two modes were distinguished" (1986:385). A multi-dimension approach is needed, which takes into account syntactic features as well as communicative functions. He then presents a method for distinguishing between text types, in which texts are classified along three scales, depending on the presence/absence of certain linguistic features. The scales are roughly supposed to say something about (1) interactive vs. edited texts, (2) abstract vs. situated content, and (3) reported vs. immediate style. This approach provides a very concrete way of distinguishing between text categories. An interesting result of this analysis is that text type is shown to be more important than the medium.

Time and space adverbials are among the features relevant for factor 2 above, since they situate the text in a spatial/ temporal context, making its reference a concrete one. On this scale the text types so far included in my material are distributed in the following way according to Biber's investigation (going from formal to informal): Press reports, broadcasts, fiction, face-to-face conversation (1986:399).

As mentioned above, the material for this paper consists of texts from the categories Conversation and Commentary. The conversation texts were chosen because spontaneous, unplanned face-to-face conversation is widely assumed to be the most frequent and most typical instance of spoken language. Commentary differs from conversation in several respects. Conversations have the form of dialogues, which means that there are several speakers who take turns and interact with each other. This also implies that speaker and hearer have the same situational frame of reference. Commentaries are mostly monologues. Although there are sometimes two or three reporters working together on a radio commentary, they very rarely interact. The speaker and the listeners do not interact at all. Besides, the medium (radio) compels the speakers not to take long breaks, i.e. they have to keep talking for as long as the programme is on the air.

The contributions to a conversation are not planned beforehand. The same is partly true for commentaries, but only to a certain extent. Although the speaker cannot predict exactly what will happen, and accordingly, what will have to be reported, the main framework is given from the outset. Most of the words and expressions will be taken from the same lexical field, and so the speaker's commentaries can be at least well prepared, although not planned in detail.

Enkvist 1982 proposes a system to show the degree of "impromptuness" of a text. The variables are (a) degree of scripting, (b) extent of planning, and (c) degree of macrostructural boundness. A text, though spoken, can be wholly or partly prepared in the form of a manuscript (variable a). Even if there is no manuscript present, it is possible that the speaker has a fairly good idea of the speech situation beforehand, so that he is well prepared, e.g. for an interview (variable b). "Degree of macrostructural boundness" says something about how stereotyped the speech situation is. Examples of bound texts are the liturgy in church, court proceedings, and greetings.

These variables bring out differences between the Conversation and the Commentary categories. Commentaries are probably not scripted, but very likely planned. There is also a certain degree of macrostructural boundness present, in that there are conventions for the form of different types of commentaries. A conversation is obviously not scripted, nor does it normally involve a lot of planning. There is a minimum of macrostructural boundness, the most important factor being the presence of at least two speakers who take turns, which distinguishes the text from a monologue.

Thematically, too, commentary is a relatively bound category as compared to conversation. In a conversation the topic is constantly being negotiated between the speakers, and the form will to some extent change according to the topic. There is no fixed set-up of a conversation, apart from what is implied in Grice's conversational maxims (cf e.g. Brown & Yule 1983:31f). With commentaries, however, listeners have quite strong expectations about what they are about to hear. The topic is given once and for all, for instance in the title of the programme, and the speaker(s) will be expected to stick to it quite strictly. In a football commentary, for example, the speaker has an obligation to report the events on the field, and not talk about other things. In case nothing exciting is going on, there seems to be a very limited set of other topics available; they are either linked with the main topic (such as information about the players), or they are informationally very light (so that remarks about the weather wil! be acceptable, but not comments on the political state of affairs).

In most cases commentaries can be said to follow a *script*. (A script "incorporates 'a standard sequence of events that describes a situation'" (Riesbeck & Schank 1978, quoted in Brown & Yule 1983:243)). In sports commentaries the sequence of events is determined to a large extent by the rules of the game. In other commentaries there may also be strong conventions which determine the sequence of events, as well as the form of the commentary, e.g. in the reports on Princess Anne's wedding, and on Sir Winston Churchill's funeral. This, of course, links up with Enkvist's concept of macrostructural boundness.

It follows from this that commentaries are relatively homogeneous, predominantly unitype texts, whereas conversations can be multitype. That is, a commentary is almost exclusively descriptive.[1] A conversation, on the other hand, can shift between various types according to the current topic; narrative, argumentative, descriptive, and perhaps even expository.

Another factor which distinguishes the two text categories is that of time. In a conversation speakers can indulge in a topic for as long as the other participants in the conversation will allow it to go on. In a commentary, on the other hand, the reporting takes places practically simultaneously with the reported action. This implies that the speaker has a limited amount of time at his disposal; it is the sequence of events that determines the pace of the commentary. It seems reasonable to see this time factor in relation to the density of information packing. If a lot of events are taking place at the same time, the information will have to be densely packed. In case very little happens, pieces of new information will be few and far between.

In view of all this, it should be clear that text types differ from one another as regards both textual form and communicative function. This is why I find it interesting to investigate how such differences affect the syntax of the texts as well.

4. Syntactic analysis of spoken English

A syntactic analysis of spoken language entails particular difficulties. The most obvious one is that of identifying a clause, a sentence, or some other unit which can be a workable point of departure for syntactic analysis. Crystal (1980:159) claims that "the *clause* is the unit in terms of which the material is most conveniently organized." According to Crystal such an analysis "correlates much better with a prosodic analysis of such data [speech], and thus with a possible model of speech production" (1980:160).

Accepting Crystal's argument, the clause, not the sentence, is the unit I work from. This does not of course solve all the problems, since speech (and indeed also writing) contains fragments which cannot readily be analysed as clauses. For instance in football commentaries there are many unattached adverbial phrases telling the public where the ball is going all the time, who passes it to whom etc:[2]

(3) Aston to GOWLING # Gowling to Tony DUNN # DUNN # down the LINE (S.10.2.855)

(4) out comes STEPNEY # and now left-FOOTED # his CLEARANCE # is AGAIN # a LONG HIGH # probing BALL # down CENTREFIELD # onto the head of – FLYNN # Flynn to BADGER # Badger on the far SIDE # (S.10.3.871)

A similar problem arises when the speaker starts on a series of adverbials and finally finds that the string has become too long, so that s/he has to make a new start. This is shown in (5). Pragmatically the first string of adverbials serves to locate temporally the event described in the main clause. Syntactically, however, its connection with the main clause is lost. In the example the second start has been italized:

(5) well ONCE # when POPPY # had stripped DOWN her engine in her CAR # and was STANDING there # and you were HELPING her and # *when Miss BLACK went past # she SAID # SORT of # I did NOTICE # POPPY # sort of taking her engine to PIECES # (S.1.12.341)*

In a study of word order, one has to work from structures above a minimum of syntactic constituents in order to make an account of word order meaningful and in order to be able to generalise. Since the object of this study is adverbials, I find it reasonable to require that there should be a verb present, finite or non-finite, which can serve as the nucleus of the structure, and which the adverbials are connected to. A very important practical reason for this is that the position of an adverbial is determined relative to the verb phrase. In the material for my investigation I have therefore excluded verbless clauses. This implies, then, that the sequences of adverbials such as those quoted in (3) – (5) have to be excluded from the study.

5. Some results of the investigation

As table 1 shows, there are differences between text types as to the proportion of adverbial sequences realised as clusters and combinations respectively. (I have included in the tables the figures for the written part of the material collected so far for comparison, although writing is not the concern of this paper.[3])

	Written		Spoken		
			Commentary		Conv.
	Press	Fiction	Sport	Non-sport[4]	
Clusters	79.7%	60.1%	63.0%	55.6%	64.1%
Combinations	20.4%	39.9%	37.0%	44.4%	35.9%
Total	100.0%	100.0%	100.0%	100.0%	100.0%
Number of sequences:	492	587	189	178	512

Table 1: The proportion of clusters and combinations in the adverbial sequences examined, distributed over different text types.

It is difficult to draw any hard and fast conclusions from the results presented in the above table. They surely go to support Biber's claim that neither speech nor writing can be treated as unified text types. All the categories prefer clusters to combinations, although to varying degrees. Non-sport commentary and Fiction are the categories which have the highest proportion of combinations. Since spoken language is the main concern of this paper, and since the Fiction material has not yet been properly examined, I shall not speculate here on the reasons for the dramatic difference between the two written categories included in the material so far, even though these preliminary results certainly raise a lot of interesting questions.

Among the spoken categories Sport commentaries and Conversation have a similar distribution of clusters and combinations. The Non-sport commentaries seem to behave somewhat differently, although the difference is not a dramatic one.

The higher proportion of combinations in the Non-sport commentaries may have to do with the pragmatic function of initially placed adverbials typical of this genre, which I will come back to below. It is a fact that initial position is exploited more often in the non-sport commentaries than in the other categories (19.5% of the adverbials are placed in initial position, as against 12.5% and 10.8% in sport commentaries and conversation respectively). Not all adverbials are possible in initial position. This is especially true of obligatory adverbials, and often those that represent new information as well. Thus if a sequence contains one such adverbial, and another that has been topicalised for pragmatic or other reasons, a combination will be the only kind of sequence acceptable.

Table 2 shows that the greater tendency for non-sport commentaries to have adverbials in initial position goes for clusters as well as for individual adverbials.

	Press	Fiction	Sport	Non-sport	Conversation
Initial	3.3	5.7	5.9	20.2	4.0
Medial	–	0.6	1.7	1.0	0.3
iE[5]	2.3	0.3	–	1.0	0.6
End	94.4	93.5	92.4	77.8	95.1
Total	100.0%	100.0%	100.0%	100.0%	100.0%
Total number of clusters	392	353	119	99	328

Table 2: The position of clusters in the different text types.

Non-sport commentaries have a much higher proportion of clusters in initial position than the other text types, which display striking similarities as to the position of clusters. It is therefore natural to take a closer look at the clusters in initial position in this category.

Of the 18 initially placed clusters in the Non-sport commentaries all except three contain at least one adverbial denoting space position. The typical function of these adverbials is to serve as a neutral background for the presentation of something new on the scene. In a way they may be said to serve as presentative constructions, similar in function to the existential *there*-construction.

(6) the four Pullman COACHES # have gone past and *NOW # IMMEDIATELY in front of the QUEEN* # comes the Royal SALOON # (S.10.7.577)

(7) and *HERE # NOW* # just framed in the DOORWAY # we see Her Royal Highness the BRIDE # (S.10.6.732)

These adverbials characteristically have a low information value, indeed they may often be given in or inferrable from the context. They provide an informationally weak opening of the sentence so as to give maximal focus to the new element that comes last. The verb is typically one of existence/appearance on the scene.

The fact that initial position can be used for informationally weak elements is sometimes carried to the extreme in examples like the following:

(8) here *at Victoria STATION* # we're awaiting a MESSAGE # (S.10.7.314)

(9 because *as I SPEAK # directly in FRONT of me # HERE* # COMES # the glittering procession of the captain's escort with STANDARD # (S.10.6.247)

(8) has been taken from a text in which everybody is waiting for the King and Queen of Nepal, who are just about to arrive, but whose plane is late. Thus, as the programme goes on the air, absolutely nothing is happening, but the reporter has to say something to fill in time. Both of the initially placed adverbials convey information which is very well known, so that they could not possibly occupy a position in which they would receive focus. The pragmatic meaning of the whole utterance is probably something like "we're still here, don't turn off the radio".

In (9) the piling up of adverbials in initial position probably serves both of these functions at the same time: that of providing a background in a presentative construction, and that of filling in time. The latter function is further evidenced by the high number of tone unit boundaries found in this utterance. It is not in the nature of a procession to move quickly, and so this slow pace is reflected in the reporter's speech, which is slowed down by means of breaks and informationally superfluous adverbials.

Some of the radio commentaries have several reporters posted in various places. In such cases initially placed markers of setting may be necessary when there is a new speaker coming in, so that the audience will know what is being reported. This shows that initial position can be used also for constituents which are crucial for the further interpretation of the text (cf. The "principle of crucial information first, Virtanen 1988).

(10) *HERE # beside the river THAMES # at Tower HILL # London City's ancient FORTRESS #* the Tower ITSELF # is the SETTING # of what is almost the FINAL STAGE # of this funeral JOURNEY # (S.10.5.1253)

In this example the reporter has just taken over for somebody else who has covered the funeral service in St Paul's cathedral. It is important then, to mark the new setting explicitly, even if nothing of what is otherwise being communicated here seems to have much information value. It is perhaps characteristic of a text where there is no need to pack information very densely that the speaker can afford to let a sentence serve the one function of marking the setting.

It is also possible that initially placed adverbials marking temporal or spatial position are typical of the reporting style. If so, what we have in cases such as (8) – (10) is a kind of phoney report, where there is little content, but where the appropriate form, or syntactic pattern, has been retained.

6. *Adverbials as pause-fillers*

It is not only in initial position that informationally and semantically light adverbials can be thrown in as "pause-fillers". This is particularly conspicuous in the clusters in which a time adverbial precedes a space adverbial, contrary to the "ordinary" pattern. The pause-filler adverbials do not occupy clause-final position, in which they would receive unintended focus. Rather they are placed in front of a semantically heavier adverbial in a cluster in end position.

(11) (the Duke of Kent accompanied by the Crown Prince) and also the Nepalese AMBASSADOR # are WAITING # *at this*

MOMENT # at GATWICK # (S.10.7.431)

(12) ... and ran well in (the) *AFTERWARDS # in the CESAREWITCH #* (S.10.4.654)

(11) has been found in the previously quoted text about the delayed plane of the King and Queen of Nepal. The adverbial "at this moment" seems to have no other function than that of filling in time; informationally it is given by default in a commentary of this kind, and semantically it adds nothing to the message.

(12) is interesting in that it shows explicitly how a semantically light adverbial is inserted in front of the more heavily weighted second adverbial. What is presumably a false start here, indicates that the speaker needs more time to remember the name of the location given in the second adverbial. (11) is from a non-sport commentary, (12) from a sport commentary. I have not found any clear examples of this pause-filler function in the conversation texts.

7. Time+Space clusters in End position

As mentioned in section 1, there is a marked tendency for spatial adverbials to precede temporal ones in clusters in end position. This tendency is clear in all the text categories examined. There is, however, a higher proportion of T+S clusters in Commentaries (Sport and Non-sport) than in Conversation. The proportion is shown in Table 3. Only clusters with 2 adverbials in End position have been included in the table.

	Conversation		Non-sport		Sport	
Space + Time	215	73.6%	24	38.1%	47	49.0%
Time + Space	16	5.5%	11	17.5%	7	7.3%
Time + Time	28	9.6%	6	9.5%	7	7.3%
Space + Space	33	11.3%	22	34.9%	35	36.5%
Total	292	100.0%	63	100.0%	96	100.0%

Table 3: the internal organisation of clusters in End position containing two space and/or time adverbials.

It is possible that the higher proportion of T+S clusters in Commentaries can be attributed to the fact that semantically, syntactically and informationally light adverbials occur quite frequently in this category. The single-word adverbials *now*, *again*, *just*, *still*, *there*, *here*, occur more than twice as frequently in Commentaries as in Conversation. These adverbials do not usually add much to the meaning of the clause. For this reason they are unlikely to receive focus, and thus to occupy clause-final position. On the other hand they can easily serve as pause-fillers, since they can be more or less given by default in the context. If such an adverbial occurs in a cluster together with a longer adverbial, it will also tend to be placed cluster-initially, due to the principle of end weight (cf. Quirk et al 1985:1361f).

In fact, a lot of the T+S clusters have a single-word temporal adverbial in cluster-initial position.

(13) the clip-CLOP # is just coming *now into our EAR* (S.1.6.595)

(14) and Carbon coming *again on the far SIDE* (S.10.4.98)

(15) but they're not active *TODAY at the lek* (S.10.8.356)

Meaning is also relevant here. Adverbials which have a meaning roughly equivalent to "now", for example, may also occupy cluster-initial position in a T+S cluster. Strictly speaking, such adverbials are informationally superfluous when the verb phrase is in the present tense. Thus they are badly suited to fill the focal end position. (16) was quoted and discussed in section 6, but is repeated here for convenience. It is the spatial adverbial which represents the new piece of information, whereas the temporal adverbial is given in the situational context.

(16) (the Duke of Kent accompanied by the Crown Prince) and also the Nepalese AMBASSADOR # are WAITING # *at this MOMENT # at GATWICK* (S.10.7.431)

8. Conclusion

A purely syntactic description may not be sufficient when dealing with the placement of constituents which are optional and whose placement is relatively free, even in a language like English which has a relatively fixed grammatical word order. Factors such as genre and medium are shown to have an influence on the shape of a text, and temporal and spatial adverbials can be used for a range of secondary functions, having to do with the total build-up of the text. This paper will have shown the importance of taking such pragmatic factors into account, in order to explain the word order variation found in the organization of sequences of temporal and spatial adverbials in English.

Notes

1. The text types have been labelled here according to the system set up in Werlich 1983.
2. The character # serves to mark tone unit boundaries. Words with nuclear stress are written in capital letters.
3. The written material has been taken from (1) British newspapers, 1988; (2) three crime novels: James, P.D., 1980: *Innocent Blood*, Penguin. Rendell, R., 1987: *A Fatal Inversion*, Penguin. Taylor, A., 1987: *Freelance Death*, Penguin.
4. I have chosen to keep the sports commentaries apart from the 'non-sport' commentaries so far in the analysis, because as the table shows, they seem to behave differently as regards adverbial sequences. Each category thus contains four 5000 word texts.
5. Initial End position, i.e. the position after the Predicator, but before an obligatory Object, Complement, or Adverbial.

References

Biber, D, 1986. "Spoken and Written Textual Dimensions in English: Resolving the Contradictory Findings." *Language*, 62:2, 384–414.

Brown, G. & G. Yule, 1983. *Discourse Analysis*. Cambridge: Cambridge University Press.

Crystal, D, 1980. "Neglected Grammatical Factors in Conversational English". In Greenbaum et al. 1980, pp 153–166. Enkvist, N.E.

& V. Kohonen (eds) 1976: *Reports on Text Linguistics: Approaches to Word Order*. Åbo: Åbo Akademi.

Enkvist, N.E. 1976a. "Prolegomena to a symposium on 'the Interaction of Parameters Affecting Word Order'". In Enkvist & Kohonen, 1976, pp 5–13.

Enkvist, N.E. 1976b. "Notes on Valency, Semantic Scope and Thematic Perspective as Parameters of Adverbial Placement in English". In Enkvist & Kohonen, 1976, pp 51–73.

Enkvist, N.E. (ed), 1982. *Impromptu Speech: A Symposium*. Åbo: Åbo Akademi.

Enkvist, N.E, 1982. "Impromptu Speech, Structure, and Process". In Enkvist (ed) 1982, pp 11–32.

Firbas, J, 1986. "On the Dynamics of Written Communication in the Light of the Theory of Functional Sentence Perspective" In Cooper, C.R. and S. Greenbaum (eds), 1986: *Studying Writing, Linguistic approaches*. Beverly Hills: Sage Publ.

Greenbaum, S. 1969. *Studies in English Adverbial Usage*. London: Longman.

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

Hasselgård, H. 1988. *On Position and Function of Spatial and Temporal Adverbials*. Unpublished thesis, University of Oslo.

Quirk, R., S. Greenbaum, G. Leech, J. Svartvik 1985. *A Comprehensive Grammar of the English Language*. London: Longman.

Virtanen, T. 1988. *Discourse Functions of Adverbial Placement in English: Clause-Initial Adverbials of Time and Place in Narratives and Procedural Place Descriptions*. Unpublished thesis, Åbo University.

Werlich, E, 1983. *A Text Grammar of English*, 2nd ed. Heidelberg: Quelle & Meyer.

–ly as Adverbial Suffix: Corpus and Elicited Material Compared[1]

Lise Opdahl
University of Bergen

1. Introduction

Most existing corpora of English texts do not contain enough material for a systematic study of low-frequency items, and corpus data must therefore often be supplemented by elicited material.[2]

This article presents some data from a project I am currently working on: an investigation of factors underlying the use/non-use of the suffix *–ly* with certain verb-modifying adverbs in English. The items studied are of relatively low frequency, and the approach mainly quantitative. The aim of the article is to show what kind of tentative conclusions can be drawn from such a material for exploratory purposes on the basis of both corpus data alone and a combination of corpus and elicited material. The material referred to in the following consists of 20 adverbial pairs selected for special scrutiny in an attempt to establish which linguistic variables (grammatical, lexico-semantic, etc.) may be at work under which circumstances, and their possible mutual dependency.

2. Corpus material

The corpus material has been taken from four different corpora, three of which consist of about one million words each, namely the LOB Corpus of written British English, the Brown Corpus of written American English, a corpus of mainly post-war machine-readable British novels (from the Oxford Text Archive – referred to below as Oxf), and the London-Lund (LLC) Corpus of spoken English, based on a version containing about half a million words.

Table 1 shows the number of occurrences of the adverbs investigated when they function as verb modifiers in the various corpora.[3] The figures are preliminary, as not all criteria for verb modification have been finally established. References to genre categories are not included.[4] As for terminology, the form with the suffix *–ly* is referred to as the plus-form, while the one without this suffix is called the minus-form.

Table 1 Number of occurrences in corpus material

Adverb	**LOB**	**Brown**	**Oxf**	**LLC**	**Total**
cheap	1	2	8	4	15
cheaply	2	3	4	1	10
Total	3	5	12	5	25
clean	1	3	3	0	7
cleanly	3	2	4	3	12
Total	4	5	7	3	19
clear	2	7	10	1	20
clearly	61	69	40	15	185
Total	63	76	50	6	205
close	65	51	89	7	212
closely	30	29	24	8	91
Total	95	80	113	15	303
dear	0	0	0	0	0
dearly	7	3	1	0	11
Total	7	3	1	0	11
deep	18	9	36	6	69
deeply	19	18	29	6	72
Total	37	27	65	12	141

direct	20	2	2	3	27
directly	58	103	46	4	211
Total	78	105	48	7	238
easy	13	7	17	1	38
easily	60	76	53	16	205
Total	73	83	70	17	243
fair	0	2	5	0	7
fairly	7	5	1	0	13
Total	7	7	6	0	20
flat	5	3	18	1	27
flatly	6	7	7	1	21
Total	11	10	25	2	48
high	36	22	75	7	140
highly	3	4	6	6	19
Total	39	26	81	13	159
loud	2	12	13	2	29
loudly	15	17	24	1	57
Total	17	29	37	3	86
low	16	20	38	7	81
lowly	0	0	0	0	0
Total	16	20	38	7	81
quick	4	6	17	4	31
quickly	150	84	190	41	465
Total	154	90	207	45	496
right	11	8	11	2	32
rightly	3	3	12	5	23
Total	14	11	23	7	55
sharp	0	0	8	0	8
sharply	36	26	65	2	129
Total	36	26	73	2	137

short	3	3	2	1	9
shortly	5	7	6	2	20
Total	8	10	8	3	29
slow	3	3	3	3	12
slowly	90	109	252	23	474
Total	93	112	255	26	486
wide	6	8	19	3	36
widely	21	13	6	9	49
Total	27	21	25	12	85
wrong	12	16	7	12	47
wrongly	4	1	0	2	7
Total	16	17	7	14	54
Sum total					
minus-forms	218	184	381	64	847
plus-forms	580	579	770	145	2074
Total	798	763	1151	209	2921
Percentages					
minus-forms	25.8	21.7	45.0	7.6	100.0
plus-forms	28.0	27.9	37.1	7.0	100.0
Total	27.3	26.1	39.4	7.2	100.0

Although there are a great many important points that such a quantitative table fails to reveal, it may nevertheless function as a diagnostic chart and thus serve as a possible starting point for further study.

The figures given in Table 1 suggest several aspects which will be further explored in other contexts. Here the following points may be noted:

(1) Total frequency of occurrence for the adverbial pairs

There is a wide span as regards the number of occurrences of these adverbial pairs, ranging from 496 (for *quick – quickly*) to 11 (for *dear – dearly*). The five most frequent adverbial pairs are *quick – quickly* (496), *slow – slowly* (486), *close – closely*

(303), *easy* – *easily* (243), and *direct* – *directly* (238). The five least frequent pairs are *dear* – *dearly* (11), *clean* – *cleanly* (19), *fair* – *fairly* (20), *cheap* – *cheaply* (25), and *short* – *shortly* (29).

(2) Distribution over the corpora

LOB and Brown, of equal size and compiled along the same lines, come out very similar, both as totals and as regards distribution over the minus- and plus-forms. The Oxford corpus, containing slightly more than one million words, has more occurrences of these adverbs, which may be expected in a corpus consisting mainly of novels with a narrative style. This corpus also has a higher percentage of minus-forms, something that may be due to the higher portion of direct speech. In the London-Lund Corpus, which is considerably smaller than the others, we find relatively few adverbs; this may also be because spoken material is likely to use fewer verb-modifying adverbs, and because this corpus contains many stock phrases.[5]

(3) Plus and minus-forms

The plus-forms are most frequent with the majority of the pairs (13 of them); the seven pairs displaying a majority of minus-forms are *cheap*, *close*, *flat*, *high*, *low*, *right*, and *wrong*. It may be noted that with some pairs the differences in frequency between the plus- and the minus-forms are not very big, e.g. *deep* (with a majority of 3 plus-forms) and *cheap* (with five more minus-forms than plus-forms). In one case, with *dear*, no minus-forms are found, while in the case of *low* no plus-forms occur.

The biggest differences between the minus and the plus-form are found with *slow* – *slowly* (12 vs. 476, i.e. more than 450 more plus-forms) and *quick* – *quickly* (31 vs. 426, i.e. almost 400 more plus-forms).

As LOB and Brown are comparable corpora, these may be contrasted for occurrences in British and American English. The distribution is shown in Tables 2 and 3 below.

Table 2 Percentages of minus and plus-forms within LOB and Brown

Form	LOB	Brown
minus-forms	27.3	24.1
plus-forms	72.7	75.9
Total	100.0	100.0

Table 3 Percentages of minus and plus-forms distributed on LOB and Brown

Form	LOB	Brown	Total
minus-forms	54.2	45.8	100.0
plus-forms	50.0	50.0	100.0

The percentages in these tables show that there is a difference in the number of minus-forms in the two corpora: LOB contains more minus-forms than Brown does. This result runs counter to the claim[6] that the minus-form is more frequently used in American than in British English.

(4) Semantic points

The antonyms included (*high – low*, *right – wrong*, *quick – slow*) display – perhaps not surprisingly – similar patterning as regards the relative frequency of plus and minus-forms.

In the following I have selected for further consideration in connection with elicited material two pairs of adverbs, namely *low – lowly*, and *direct – directly*. The figures in Table 1 indicate rather different characteristics of these two pairs: *low – lowly* is the only adverbial pair displaying no examples of the plus-form, while *direct – directly* is the pair displaying the biggest difference between LOB and Brown in the number of minus-forms (only 9.1% of the minus-forms are found in Brown).

The elicited material will now be considered in order to see whether the tendencies found in the corpus data for these two adverbial pairs can be said to be confirmed there.

3. *Elicited material*

The elicited material was obtained through a questionnaire, consisting of some 180 items in judgment test form, given to 100 native speakers of English, 50 British and 50 American, of both sexes and of varying ages and educational backgrounds, thus yielding about 18,000 answers. The respondents to the questionnaire were selected rather randomly among native speakers of British or American English who happened to be available. However, in order to make the outcome of the study as relevant as possible also for practical teaching purposes, it was decided to focus on informants with relatively high education (mostly university level), who might be expected to use some kind of 'educated' English.

In view of the relatively few occurrences of the forms in question in the corpus data I decided to try to validate these as much as possible through the material selected for elicitation. However, in order to give this stage of the study a wider scope, two other types of material were occasionally included, namely sentences to test statements by other scholars, and items intended specifically to pinpoint possible variables at work.

With this somewhat mixed basis for the setting up of the questionnaire many methodological problems were keenly felt, particularly as over the last 20 years the methods of elicitation have been considerably refined.[7] This has led to much more sophistication as regards e.g. representativeness of material and factors in the experimental situation that may influence informant responses. However, for practical reasons in this initial exploratory survey a rather crude test design was decided on.

Another limitation was that with the overview purpose few test sentences could be included of each adverbial pair: an average of 9 sentences per pair (some 180 questionnaire items distributed over 20 adverbs) clearly cannot capture many of the variables that might be at work in this rather complicated area.

Thus the results presented here cannot offer conclusive evidence, but may be suggestive for further research.

The elicited material discussed in the following is restricted to preference tests.

4. *Corpus and elicted material compared:* *low – lowly and direct – directly*

4.1 Low – lowly

Sentences given

The following five sentences were given in order to test attitudes to the use of *low* or *lowly*, here listed in the order of appearance in the questionnaire:

(45) That college was rated *low/lowly* for decades.

(81) This really made her think *low/lowly* of him.

(153) The bird was flying *low/lowly* over the lake.

(165) She saw the bird fly *low/lowly* over the lake.

(169) He dropped it *low/lowly* and carefully.

Basis for selection of test sentences

As all examples in the corpora occurred with the minus-form, the aim of the elicitation was partly to test whether this feature could be said to be representative of the usage, partly to see under what conditions a plus-form could be prompted. For the questionnaire three sentences were selected with a corpus basis (nos. 45, 153, and 165). In order to test Poutsma's statement that '*low*, when denoting place, never takes the suffix *–ly*'[8] – which may be interpreted to mean that with a figurative meaning the plus-form is more likely – sentences with different kinds of verbs were selected: nos. (153) and (165) with the verb *fly*, suggesting a literal sense of *low/lowly*, and sentence (45) with the verb *rate*, yielding a rather figurative sense.[9] Also included to prompt the plus-form was sentence (81) with *think* giving a figurative meaning; in Poutsma this verb occurs in an example with the plus-form (although there coordinated with another sentence containing the adverb *highly*).[10] Yet another way of trying to prompt the plus-form was to coordinate *low/lowly* with another adverb ending in *–ly* (no. 169). However, as one might possibly imagine a preference of the minus-form for euphonic reasons in such a case, this point will have to be further investigated. It was also thought sensible to test whether the length of the verb – in this case the progressive versus the simple form – would influence the attitude (nos. 153 and 165).

Responses

The responses to the sentences containing *low/lowly* are given below in Table 4.

Table 4 Responses on low/lowly

Sentence		only M	only P	both	reject	M	P
45	BrE	40	6	3	1	43	9
	AmE	49	0	1	0	50	1
	Total	89	6	4	1	93	10
81	BrE	11	21	3	15	14	24
	AmE	14	25	4	7	18	29
	Total	25	46	7	22	32	53
153	BrE	50	0	0	0	50	0
	AmE	50	0	0	0	50	0
	Total	100	0	0	0	100	0
165	BrE	48	1	1	0	49	2
	AmE	50	0	0	0	50	0
	Total	98	1	1	0	99	2
169	BrE	25	4	1	20	26	5
	AmE	27	13	2	8	29	15
	Total	52	17	3	28	55	20
Sum total							
	BrE	174	32	8	36	182	40
	AmE	190	38	7	15	197	45
	Total	364	70	15	51	379	85

Percentages						
BrE	69.6	12.8	3.2	14.4	72.8	16.0
AmE	76.0	15.2	2.8	6.0	78.8	18.0
Total	72.8	14.0	3.0	10.2	75.8	17.0

Discussion

Table 4 shows that the minus-form is strongly preferred in three of the four corpus-based sentences, namely in nos. (45), (153), and (165). The unanimous response in the case of (153) may be noted: of the some 180 items in my questionnaire only three others got this kind of reaction. Sentence (169), where another adverb with the suffix *-ly* was added, is acceptable with the plus-form to more respondents, but here the high number of rejections may be noted – evidently this 'doctoring' of the sentence was not judged to be very successful by the native speakers. Sentence (81), with a majority of responses in favour of the plus-form, to a certain extent confirms Poutsma's example with the plus-form, even without the parallelism with *highly*. On the basis of these responses it is, however, difficult to draw any definite conclusions as regards the influence of the distinction literal – figurative; with the verb *think* a figurative meaning is used, but also *rate* in sentence (45) may be considered figurative in this context, although perhaps not to such an extent as *think*. Here it may be noted that the American respondents strongly favour the minus-form, in a sentence taken from the Brown Corpus, while about 20 per cent of the British informants accept the plus-form. The responses to sentences (153), (165), and also to a certain extent (169), seem to support the use of the minus-form when a literal sense is implied.

Conclusions

On the basis of the corpus and the elicited material several points may be made about the use of the plus- and the minus-forms of *low*, some of which are:

(1) The situation indicated by the corpus results – that the plus-form is not frequent – is also signalled by the elicited material.

One exception indicated by the elicited material is with modification of the verb *think*. In some cases, for example with modification of the verb *fly*, the minus-form is very firmly established.

(2) The distinction between figurative and literal meaning does not seem to have enough explanatory power for the choice between the minus- and the plus-form – cf. the verb *rate* modified by the minus-form in the (Brown) corpus and a relatively strong preference for the same form by the informants.

Obviously more research on these and other points is necessary.

4.2 Direct – directly[11]

Sentences given

The following 11 sentences were given as preference tests with *direct/directly*, here presented with the original numbering, but grouped so that similar sentences occur together.

(13) This is a system that works *direct/directly* against the objective.

(113) This is a team that works *direct/directly* from nature.

(44) They decided to apply it *direct/directly* to the affected skin.

(124) They decided to apply it slowly and *direct/directly* to the affected skin.

(34) They decided to appeal *direct/directly* to the government.

(158) They decided to appeal *direct/directly* to the government this case which has troubled them for so long.

(48) They decided to explain it *direct/directly* to the young people.

(151) They decided to purchase it *direct/directly* from the overseas market.

(84) They decided to purchase *direct/directly* from the overseas markets the commodities that they needed.

(71) The committee should have the right to deal *direct/directly* with the appropriate office.

(170) The individual should have the right to deal *direct/directly* with his employers.

Basis for selection

The major basis for selection from the corpus material was which lexical verb was modified.

Six verbs were selected from the corpus material, namely *appeal, apply, deal, explain, purchase,* and *work*. Of these verbs two, namely *appeal* and *purchase*, occur only with the minus-form, one occurs only with the plus-form, namely *explain*, while the remaining three verbs, *apply, deal,* and *work*, occur with both the minus- and the plus-form.

Responses

The responses to these sentences are presented in Table 5 below.

Table 5 *Responses to sentences with direct/directly*

Sentence		only M	only P	both	reject	M	P
13	BrE	0	50	0	0	1	50
	AmE	0	49	1	0	1	50
	Total	0	99	1	0	1	100
113	BrE	6	29	15	0	21	44
	AmE	3	38	8	1	11	46
	Total	9	67	23	1	32	90
44	BrE	10	21	18	1	28	39
	AmE	3	41	6	0	9	47
	Total	13	62	24	1	37	84
124	BrE	5	39	6	0	11	45
	AmE	2	45	3	0	5	48
	Total	7	84	9	0	16	93
34	BrE	7	22	21	0	28	43
	AmE	0	44	6	0	6	50
	Total	7	66	27	0	34	93
158	BrE	10	18	21	1	31	39
	AmE	1	36	12	1	13	48
	Total	11	54	33	2	44	87

48	BrE	9	29	12	0	21	41
	AmE	1	46	3	0	4	49
	Total	10	75	15	0	25	90
151	BrE	14	15	21	0	35	36
	AmE	2	30	18	0	20	48
	Total	16	45	39	0	55	84
84	BrE	20	14	16	0	36	30
	AmE	4	29	17	0	21	46
	Total	24	43	33	0	57	76
71	BrE	10	20	20	0	30	40
	AmE	0	43	7	0	7	50
	Total	10	63	27	0	37	90
170	BrE	9	20	21	0	30	41
	AmE	0	42	8	0	8	52
	Total	9	62	29	0	38	91
Total							
	BrE	100	277	171	2	271	448
	AmE	16	443	89	2	105	532
	Total	116	720	260	4	376	980
Percentages							
	BrE	18.2	50.4	31.1	0.4	37.7	62.3
	AmE	2.9	80.5	16.2	0.4	16.5	83.5
	Total	10.6	65.5	23.6	0.4	27.7	72.3

Discussion

One of the points shown by this table is that with one exception – sentence (13), where the reactions are equal – the American respondents systematically favour the plus-form to a considerably greater extent than do the British ones. This preference for the plus-form among the Americans with this adverbial pair might have been even more difficult to interpret if the corpus data had not been taken into consideration.

Many scholars[12] suggest that *direct* and *directly* overlap only partly in meaning or reference, stating that the minus-form generally has a spatial function, while the plus-form may have both a spatial and a temporal use. With this in mind the elicited material standing alone might be interpreted as an expression of a much stronger tendency of the Americans to see these sentences as ambiguous – having both a spatial and temporal meaning. However, a juxtaposition with the corpus material shows that a more likely interpretation is that the minus-form by many American speakers is not generally used as a verb modifier (see only 2 occurrences in the Brown Corpus), or if used, is limited to very special contexts or styles, and that the plus-form functions in a satisfactory way for these speakers in all the relevant spatio-temporal aspects. This may again be supported by the figures in the elicited material, where we find a total of not more than 16 options for *direct* as the only form out of the 550 American responses, while the corresponding British number is 100.

More comments could be made on this material, but I hope to have given another example of the usefulness of a comparison of the corpus and elicited material.

4. Concluding remarks

In this article I have tried to show that both corpus data alone and juxtaposition of corpus and elicited material can be used for exploratory purposes in the study of low-frequency items.

A direct link between corpus and elicited material is not always easy to establish. However, in the same way as certain guidelines have been developed for establishing representative corpora, the link between corpus and elicited material may be strengthened by attempting to develop guidelines for the sampling from corpora of sentences for elicitation. This may not be an easy task. A quotation from Greenbaum – although not originally used with reference to corpus studies – may serve as a suitable starting-point for such a venture. He states (1977a:6): 'just as there is a scale of acceptability, so there is a scale of frequency in use; and the two scales do not necessarily coincide, though it is reasonable to expect some relation between them.' This relation should be further explored.

Notes

1. An earlier version of this article was read as a paper at the 10th ICAME Conference in Bergen, June 1–4, 1989.
2. Probably the best-known example is the Survey of English Usage, the sources of which have been presented in many publications, one of them being Greenbaum 1984:197.
3. For the LOB and Brown corpora the frequencis of the graphic words are found in Hofland and Johansson 1982. In these corpora the words in question may be identified as adverbs through the tags now added. However, as this tagging does not display the function of these words as verb modifiers, the figures given here are based on my own manual syntactic classification.
4. The main reason is the relatively few total occurrences of the forms in question, which may make their occurrences rather accidental. However, the adequacy of the genre categories in LOB and Brown has also been questioned (see e.g. Biber and Finegan 1986 and Oostdijk 1988).
5. See Altenberg 1989.
6. See e.g. Kirchner 1979:233.
7. Important contributions have been made first and foremost by Greenbaum in a series of publications from the late 60s onwards (e.g. Greenbaum 1969, 1970, 1973, 1977a, 1977b, 1984), but also by other scholars, like Quirk (e.g. Greenbaum and Quirk 1970, Quirk and Rusiecki 1982, Quirk and Svartvik 1966) and Svartvik (1968) in Europe and Bolinger (1968) in America.
8. See Poutsma 1926, 2,2:628.
9. Corpus sentences with the verb *lie* were avoided, as this vberb is often included in lists of verbs that may have a copula-like

function (see e.g. Poutsma 1928 1,1:10 and Quirk et al. 1985:1168 and 1172).

10. Poutsma's (1926, 2,2:628) example with *lowly* occurs in the following sentence: 'He had thought highly of Desert; and – odd! – he still did not think lowly of him.' (taken from Galsworthy, *The White Monkey*, I, ch. 9,76). In this sentence *lowly* is paralleled with *highly*, which may have influenced the form chosen.

11. The data presented for this adverbial pair are discussed more extensively in Opdahl 1989.

12. E.g. Fowler (1965:133), Weiner (1983:102), and Quirk et al. (1985:407).

References

Altenberg, B. 1989. Collocations in Spoken English. Paper read at the 4th Nordic Conference for English Studies, Elsinore, May 11–13, 1989.

Biber, D. and E. Finegan 1986. An Initial Typology of English Text Types. In Aarts, J. and W. Meijs (eds.) *Corpus Linguistics II. New Studies in the Analysis and Exploitation of Computer Corpora*. Amsterdam: Rodopi. 19–46.

Bolinger, D. 1968. Judgments of Grammaticality. *Lingua* 21, 34–40.

Fowler, H.W. 1965. *A Dictionary of Modern English Usage*. 2nd ed. Oxford: Clarendon Press.

Greenbaum, S. 1969. *Studies in English Adverbial Usage*. London: Longman.

Greenbaum, S. 1970. *Verb-Intensifier Collocations in English: An Experimental Approach*. The Hague: Mouton.

Greenbaum, S. 1973. Informant Elicitation of Data on Syntactic Variation. *Lingua* 31, 201–12.

Greenbaum, S. 1977a. Contextual Influence on Acceptability Judgments. *International Journal of Psycholinguistics* 6, 5–11.

Greenbaum, S. 1977b. The Linguist as Experimenter. In Eckman, F. (ed.) *Current Themes in Linguistics: Bilingualism, Experimental*

Linguistics and Language Typologies. New York: Wiley. 125–144.

Greenbaum, S. 1984. Corpus Analysis and Elicitation Tests. In Aarts, J. and W. Meijs (eds.) *Corpus Linguistics. Recent Developments in the Use of Computer Corpora in English Language Research.* Amsterdam: Rodopi.

Greenbaum, S. and R. Quirk 1970. *Elicitation Experiments in English: Linguistic Studies in Use and Attitude*. London: Longman.

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

Kirchner, G. 1970. *Die Syntaktischen Eigentümlichkeiten des Amerikanischen English* 1. Munich: Max Hueber.

Oostdijk, N. 1988. A Corpus Linguistic Approach to Linguistic Variation. *Literary and Linguistic Computing* 3, 12–25.

Opdahl, L. 1989. 'Did They Purchase it Direct – or Directly?' On *Direct* and *Directly* as Verb Modifiers in Present-Day British and American English. In Breivik, L.E., A. Hille, and S. Johansson (eds.) *Essays on English Language in Honour of Bertil Sundby*, Studia Anglistica Norvegica 4. Oslo: Novus. 245–257.

Poutsma, H. 1926–29. *A Grammar of Late Modern English*. Groningen: Noordhoff.

Quirk, R., Greenbaum, S., Leech, G., and Svartvik, J. 1985. *A Comprehensive Grammar of the English Language*. London: Longman.

Quirk, R. and J. Rusiecki 1982. Grammatical Data by Elicitation. In Anderson, J. (ed.) *Language Form and Linguistic Variation*. Amsterdam: Benjamins. 370–394.

Quirk, R. and J. Svartvik 1966. *Investigating Linguistic Acceptability*. The Hague: Mouton.

Svartvik, J. 1968. Plotting Divided Usage with 'Dare' and 'Need'. *Studia Neophilologica* 40, 130–140.

Weiner, E.S.C. 1983. *The Oxford Guide to English Usage*. Oxford: Clarendon Press.

The Australian Corpus Project and Australian English

Elizabeth Green and Pam Peters
Macquarie University

Plans for compiling a corpus of Australian English along the same lines as the Brown and LOB corpora were described in *ICAME Journal* No.11 (Peters, 1987), and have since been put into action at Macquarie University. The scope for three-way comparisons is particularly useful in profiling the character of contemporary Australian English, seeing that Australia's relationships with Britain and USA have changed radically since World War II. The strengthening of strategic and cultural links with North America has resulted in many lexical loans,[1] which are used enthusiastically by some Australians and are a source of indignation for others.

Yet it is easy to exaggerate the American influence on Australian English, and we wanted to see whether there were signs of the American connection in the less conspicuous elements of our local variety, in morphology, the use of function words and routine details of grammar. The Americanisation of such details would in fact indicate that the American influence was much more profound than we might have realised, since there are no strong cultural pressures to prefer *maybe* to *perhaps* or *toward* over *towards*. Our interest in the more functional elements of Australian English was also something which could reasonably be researched through a data base of one million words, which is of course too small for middle to low frequency lexical items including neologisms. With a corpus of Australian English which matched both Brown and

LOB, we could make quantifiable comparisons on the grammatical idiom of all three varieties.[2]

Apart from facilitating intercomparisons, the Brown/LOB model suggested a data base of written genres which are functionally similar in their relationship to Australian English, as those of the American and British corpora were to their respective varieties. In each the corpus embodies a wide spectrum of published English which is freely accessed by local readers, and composed by local writers from many quarters of society. The contents of the corpus are thus more broadly representative of local written usage than in societies where English is a second language (cf. Leitner, 1991), or where the writers of English are a subset of the total literate community.

In compiling a corpus of this type we were under no illusions that it would provide insights into spoken usage, or any kind of situational language which does not find its way into print. A million-word corpus cannot do justice to the full range of situational and sociolinguistic variables. Yet the Brown/LOB genres do represent a number of the variables of written usage, including both fiction and nonfiction, popular versus scholarly and literary writing, material for general and for specialised readerships, and serial publications versus monographs. Those several parameters ensure that the contents of the corpus are quite heterogeneous, and that the chances of word frequencies being skewed by too many texts from too few sources is reduced. Heterogeneity was also important to us because the Macquarie Corpus is the first systematically compiled corpus of Australian English, and one which we hope will serve as a reference point for more specialised Australian corpora in the future.

Although there were a number of reasons for compiling our corpus according to the Brown/LOB model, it proved difficult in several categories to match their samples with equivalent Australian material. Some of these difficulties reflect differences in geographical location and the small size of our population; others may reflect common changes in social concerns and fresh intellectual emphases which have developed everywhere since the Brown Corpus was begun in the 1960s. At any rate these problems in sampling affected both the types and relative quantities of texts being published and available for sampling in Australia in 1986.[3] The dearth of material forced us to adjust the proportions of texts within certain categories, e.g. the various types of press reportage in Category A (Collins and Peters, 1988); and to modify the balance that was established in Brown and LOB between monographs and articles in journals when it came to Category J, learned and scientific writing. But the greatest divergence from the original sampling model was in the compilation of the fiction categories K – R, where several kinds of writing proved quite elusive in Australia (see below). Categories D – H meanwhile posed no such problems, there being an ample supply of general expository writing on topics from religion to racing, as well as institutional reports and documents.

The need to find substitutes raised questions about the original motives for sampling this and that: Were the deciding factors (i) subject matter, (ii) style or level of writing, (iii) medium of publication, (iv) relative influence or authority, (v) size of print run or circulation? The first two of those are the essential parameters of genre, and they seem to have been regularly involved in selecting samples for Brown and LOB.[4] The third is built into the practice of including extracts from both monographs and periodicals in categories D to J, and from books as well as short story anthologies in K – R. Like the first two factors, the third contributes to the generic identity of a source text; though it also affects the perceived authority and circulation of the text (factors four and five). The latter is its "reception index", to borrow Nelson Francis's expression (1982). The reception index was more deliberately involved in some categories (notably A to C), in the weighting given in LOB to the quality press, and to the higher circulating national papers as opposed to regional ones (*Manual*, 1978:14–16). The appeal to the reception index helps to ensure that the texts sampled are not just generically appropriate, but significant examples in terms of their consumption within a given culture. They are thus more thoroughly representative of current usage in a particular society. The representativeness of the samples in this sense is of course different from how representative they are in reflecting the range of publishing actually available in different countries. But matters like these raise enormous questions of comparability. Just how equivalent is the publishing scene and the reception index for particular kinds of publication in USA, Britain and other English-speaking countries? We will take up the question of representativeness later in the paper, and concentrate for the moment on that of generic comparability.

For the compilers of the LOB Corpus, its comparability with Brown rests on generic similarity, though at a general level rather than in the particulars of each category: "The matching between the two corpora is in terms of general categories only..." (*Manual*, 1978:4). They acknowledged divergences from Brown within categories, in the selection of newspapers (where both circulation and "quality"/relative authority were allowed to influence the sampling), and in divergent allocation of a few samples in categories E, F and G. That group is the focus of the comment that they should be seen as a composite set of "general expository writing" rather than separate categories. The builders of LOB also noted some minor problems in deciding how to match samples in Brown when it came to the boundaries between category J, and some items in both D and G. Yet such differences have not been felt to undermine the comparability of the LOB genres with those of Brown, and data from the two corpora has been the basis of numerous comparative studies, including several by Biber, Johansson, Leech, Meijs, and others.[5]

In compiling the Australian corpus, we have sought to match the internal structure of the original Brown/LOB categories as far as possible in substance as well as style. When this proved impossible we included what could at least be claimed to be generically comparable. Our Category J, for example, is consistent in the scholarly style which pervades all the articles included, but we had some difficulties there in matching the media of publication, and particular subjects. To take the first point: it soon became apparent that we would not be able to find many Australian monographs in the natural sciences, and that the Brown/LOB balance between monographs and journal selections could not be maintained. An exhaustive check of the Australian National Bibliography, supplemented by the NSW State Library (a deposit library) and the computer catalogues of major university libraries, could produce no more than two monograph titles in the natural sciences in 1986. Various textbooks for teaching science subjects were published in that year, but the small population and even smaller academic market seems to put economic constraints on specialised science publishing. With only two monographs in this area, we were forced to rely heavily on samples in journals. By contrast we found a shortage of Australian journals in applied science and technology, and in the humanities. So in these subcategories we used a much larger proportion of monographs than had been the case in Brown and LOB. Overall, we were seriously restricted in our choices between monographs and journals by what Australian publishers produce.

Table 1 *Ratio of monographs to journal articles in subcategories of category J, compared in Brown, LOB and Macquarie*

Subcategory		**Brown**	**LOB**	**Macquarie**
Natural Sciences	monographs	6	4	2
	articles	6	8	10
Medicine	monographs	2	2	2
	articles	3	3	3
Mathematics	monographs	2	1	2
	articles	2	3	2
Technology and Engineering	monographs	6	4	10
	articles	6	8	2
Social and Behavioural Sciences	monographs	8	5	9
	articles	6	9	5
Political Science, Law, Education	monographs	8	7	9
	articles	7	8	6

Humanities	monographs	10	11	13
	articles	8	7	5
TOTAL	monographs	42	34	47
	articles	38	46	33

As those totals[6] show, there are differences between Brown and LOB in the proportion of monographs and journal articles selected. In fact the LOB sampling diverges from that of Brown by almost as much as Macquarie was obliged to. The presence of more journal articles in LOB might have increased the amount of technical vocabulary a little (presuming that journal articles are addressed to slightly more specialised audiences than monographs). But with extensive sampling from both media in each corpus, any potential language differences are assumed to be balanced out. Frequencies from parallel categories in Brown and LOB, including Category J, have certainly been put to comparative use (see for example Krogvig and Johansson, 1984).

Attempts to find matching samples in Australia for the particular science subjects in Brown and LOB also proved difficult sometimes. In the natural sciences, for example, texts from the fields of rheology, radiation chemistry, cryogenics, and meteorology could not be found, and there seemed to be only one source of physics texts. On the other hand there was a relatively large number of zoological, botanical and ecological journals, demonstrating a preoccupation with environmental sciences within the Australian scientific community. This we felt should be represented among the samples of science writing, whether it reflects particular local concerns of Australian society, or a general extension of scientific interests around the world in the last twenty years.

Table 2 *The fields sampled in the natural sciences, compared in Brown, LOB and Macquarie.*

Brown	**LOB**	**Macquarie**
geology	geology	geology
rheology	rheology	
		geophysics

radiation chemistry	radiation chemistry	
chemistry		chemistry
		organic chemistry
physical chemistry		physical chemistry
physics	physics	physics
	applied physics	
	sub-atomic physics	
astronomy		astronomy
	meteorology	
biology	biology	biology
	marine biology	
		zoology
		botany
		ecology
	cryogenics	
		metascience

Australian scholarly writing on geophysics and metascience suggested new interdisciplinary kinds of science, which need to be represented, whether they are local or global developments in scientific thinking. In technology we were unable to represent areas such as textiles and metallurgy, or the exact range of engineering which LOB contains. We adjusted the sampling to include available texts in the areas of information technology, environmental technology, urban planning, and space engineering. Texts from these fields were not included in Brown and LOB, and probably represent global technological developments since the 1960s. The conspicuous interest in agricultural technology seemed to reflect local Australian geographical factors.

Table 3 *Comparison of technological and engineering fields represented in Brown, LOB and Macquarie*

Brown	LOB	Macquarie
	metallurgy	
textiles		
food technology		food technology
plastics		plastics

optics		optics
		information technology
		environmental tech.
		agricultural tech.
		urban planning
electronics		
electrical engineering		electrical engineering
naval engineering		
marine engineering		
	chemical engineering	chemical engineering
	mine engineering	mine engineering
	civil engineering	
	aircraft engineering	aircraft engineering
		space engineering

Our sampling in the humanities was seriously constrained by gaps in availability of certain subjects. For instance there seemed to be little critical writing on philosophy or art, and no vehicle for academic music criticism (music periodicals seem to function as newsletters). There were however many historical texts, and several journals of literary criticism. A noteworthy addition to the Australian list is semiotics, a new metadiscipline which has its practitioners in Australia as elsewhere.

Table 4 *Comparison of fields in the humanities, as sampled in Brown, LOB and Macquarie*

Brown	**LOB**	**Macquarie**
philosophy	philosophy	philosophy
history	history	history
art criticism	art criticism	art crit.
literary criticism	literary criticism	literary criticism
	craft criticism	
architecture		architecture
music criticism		
		semiotics

For all the differences in subject matter which we have highlighted in tables 2, 3 and 4, there are plenty of samples in the Macquarie Corpus whose subject matter matches that of samples in either Brown or LOB. The tables also serve to confirm that the matching between Brown and LOB was generic rather than at the level of individual samples. As the LOB Manual notes (p.4): "There is no one-to-one correspondence between samples, although the general arrangement of subcategories has been followed wherever possible." In Macquarie too we have endeavored to include samples from corresponding fields, but more importantly, selected those which addressed a learned or scholarly reader with their concerns, and were thus stylistically comparable with their precedents.

The most substantial rethinking and remodeling of individual categories has taken place in the sampling of Australian fiction. In undertaking the sampling, we drew on all the public bibliographical aids mentioned above, as well as commercial listings of fiction which are included in *The Australian Bookseller and Publisher*, published each month by D.W. Thorpe, Melbourne, and in the associated yearbook *Australian Books in Print*. The categories of fiction used in such publications are however very broad, and more insights into the range of the contemporary Australian fiction publishing were gained through independent surveys by Gelder and Salzman (1989), and Daniels (1989). These sources confirmed that not only were some of the Brown/LOB categories completely unavailable, but that other categories needed major redefinition, and new ones needed to be proposed. The total amount of fiction published in a year in Australia is relatively small, although it has increased considerably over the last two decades (Gelder and Salzman 1989: 1–10) There are however a number of literary magazines publishing short stories. By balancing the number of novels with an almost equal number of short stories, we were able to match in number the samples collected in Brown and LOB, though it involved sampling nearly the entire range of fiction novels published in Australia in 1986. In statistical terms this means that our sampling of this genre of publication could hardly be more representative. But it also meant that we could not omit texts if they did not fit the original fiction categories. Rather, we had to reinterpret those categories and even add to them.

Table 5 *Comparison of fiction categories in Brown, LOB and Macquarie, and the ratio of to short stories in each categories*

Category		**Brown**	**LOB**	**Macquarie**
K General fiction		20	20	9
	Short St.	9	9	20
	Total	29	29	29
L Mystery/Detective		20	21	10
	Short St.	4	3	5
	Total	24	24	15
M Science fiction	Monographs	3	3	2
	Short St.	3	3	5
	Total	6	6	7
N Adventure/Western	Monographs	15	15	4
(Bush)	Short St.	14	14	4
	Total	29	29	8
P Romance/Love	Monographs	14	16	6
	Short St.	15	13	9
	Total	29	29	15
R Humor	Monographs	3	3	10
	Short St.	6	6	5
	Total	9	9	15
S Historical fiction	Monographs			15
	Short St.			7
	Total			22
W Women's fiction	Monographs			8
	Short St.			7
	Total			15
TOTALS:	Monographs	75	78	64
	Short St.	51	48	62

[For discussion of categories N, S and W, see below]

The category of General Fiction (K) in Brown and LOB is used to accommodate a kaleidoscope of fiction writing which did not fit easily into any of the other categories. It is usually fiction which depends on something other than its plot or action for interest, often with a thematic line as its essence. Among Australian fiction titles, the candidates for Category K were often more literary in their aspirations than those which were readily grouped elsewhere. Literary magazines proved a rich source of texts for this category in Australia, and compensated for the relatively small number of monographs.

For the commercial fiction categories of Brown/LOB, it proved difficult to find enough samples which were both authored by Australians and published in Australia. Crime and detective fiction which met those criteria was relatively scarce in Australia in 1986, though it has burgeoned since then (Knight, 1990), with new authors, especially women (Jennifer Rowe, Marele Day), establishing themselves in the field. "Western" fiction was entirely absent, though there was a significant number of texts concerned with adventure in the Australian bush which seemed suitable substitutes. Australian bush narratives set their action in a primitive social setting which is the counterpart of the American frontier, where human beings are scattered and civilised behavior is not to be taken for granted. We therefore renamed Category N as Adventure and Bush, and assigned the available samples to it. Even so the numbers fell short of those in Category N of Brown and LOB.

Another curiously elusive category of fiction was Romance and Love, Category P. Inquiry into this showed that Australian romance writers send their manuscripts overseas to the pulp fiction publishers of Britain and USA. Romantic short stories published in Australian magazines were found to be syndicated, or mostly written by non-Australians. In the absence of the typical publications for this category, we used it to accommodate writing which deals with human fantasies (thus *romance* in its older sense of "literary escapism"); and writing about human passions and other emotional relationships, even though they were examined in an unromantic way. Still we were unable to match the number of samples in

Brown and LOB. Interestingly Categories P and N also proved difficult to fill for the compilers of the corpus of Indian English at Kolhapur, as comparison of the proposed and actual inventories of samples shows (see Shastri, 1985 and 1988). Western and romance publications are certainly not passé as genres of reading in Australia (newsagents have continuous supplies); but the market for them depends on economies of scale, and on vast distribution networks which can only be managed by international publishing conglomerates based in Britain and USA.

The shortfall in Categories N and P meant that both the Australian and Indian corpora had to compensate for it in other areas of fiction. In the Macquarie Corpus we compensated in several ways, one of which was to include a larger number of texts in Category R, Humor, extending it from 9 to 15. It is a rich field of publishing in Australia, hence the very satisfactory number of monographs from which samples could be taken, as well as short stories. The local or regional character of humor is often commented on, and this is clearly one area of publishing where Australia supplies its own market. The Kolhapur Corpus compensates for shortages in N and P by enlarging the category of general fiction (K) from 29 to 58 samples. In the Macquarie Corpus we preferred not to submerge the compensating samples in with all the other kinds of fiction in Category K, but to identify them in their own terms. A quite substantial number of Australian fiction monographs of 1986 lend themselves to the heading "historical fiction". They are not historical novels in the sense of contrived period pieces, but fictionalised biography from pioneering days, or sagas of immigrant families. Fiction of this kind (in our Category S) seems to reflect the now widespread interest in Australian social history, just as the other distinctive type of fiction evident in Australia in 1986 correlates with the now influential women's movement (our Category W). Candidates for this category are not simply works written by women, but ones whose fictional world projects feminist styles of thinking (Gelder and Salzman, 1990:ch.4). They foreground women's roles and sexuality, and project human relationships in a very different way from the samples in Category P, Romance and Love. One can scarcely disregard it as a category of fiction, given the significant number of titles produced in Australia in a single year; and its importance in other parts of the world is confirmed by the fact that it is the one category of fiction (feminist writing) to be added in to the corpus of Canadian English, which otherwise adheres to the fiction categories established by Brown and LOB.

Our confrontation with new kinds of fiction showed yet another field in which there is a growing amount of publishing, that of the immigrant and Aboriginal writer, either bilingual speakers of English or speakers of English as second language. The raised profile of migrant and multicultural writing is however a phenomenon of the late 1980s, and there were insufficient examples in 1986 to justify including them as a new category, apart from the very fundamental question they raise about what constitutes an Australian author. Does s/he have to be native-born? Australian publishing is now beginning to recognise cultural diversity here; and it is important for research into Australian English to consider whether it is the property of "true blue" Australians only (i.e. native-born), or a commodity shared by all those with extended residence in the country. In constructing the Macquarie Corpus we have tried to restrict our sampling to work produced by native-born Australians (using whatever biographical information was available), because this was one of the parameters established at the start of the project. But this means we have assembled a data base which is perhaps more homogeneous than it should naturally be, if we are to represent the increasingly multicultural nature of Australian society, and the heterogeneity of Australian English. In future we must ask whether it is desirable to exclude immigrant or Aboriginal writers of Australian English from a corpus of contemporary writing.

These questions of sampling return us to the more general questions of comparability and representativeness raised at the start of the paper. The Australian data base is generically comparable to Brown and LOB, in that the material collected has been carefully matched in terms of style, substance and medium wherever possible. Where there are differences in subject matter and the medium of publication, broader generic comparability is maintained through the set of categories (as with the fiction categories K – W), and within a single wide-ranging category such as J. But the search for equivalent samples has made us acutely aware of what sectors of Australian usage we might or might not be representing. For one thing, our corpus, like Brown and LOB, represents published usage rather than written usage at large. Infinite varieties of private and in-house writing are unrepresented, though they certainly do not have the circulatory reach of published material. In collecting what is published we are going for writing whose influence must be greater and more ubiquitous. And yet even among published texts, the amount of influence which each wields is uneven, a factor which is disregarded in purely random methods of sampling from bibliographical lists. We felt it undesirable to extract samples from publications which, though entered in the National Bibliography, are held in only one or two Australian libraries. While their generic suitability is unquestioned, their reception index is such that they are unlikely to have much impact on Australian usage. If the underlying aim of the corpus is to embody and represent the mainstream of current usage, more systematic attention to the reception index of source materials is needed.

Establishing the reception indices of individual publications is a challenge to researchers anywhere, because of the difficulties in showing how much any publication is read. Circulation figures can be obtained for newspapers and magazines, compiled in Australia by the Audit Bureau of Circulation, but they do not show how many people read the same copy of a newspaper or magazine, and how many copies are pulped. The sizes of print runs for monographs are not systematically collected, and publishers are reluctant to reveal them, except for very successful items. However some light has been shed on the consumption of books in Australia through recent research commissioned by the Australia Council. It surveyed both book-buyers and library users about their purchase and use of books, and provided a wealth of information about Australian reading habits. One interesting finding was the consistently high level of consumption of Australian fiction, on a par with that of all varieties of Australian nonfiction taken together (Guldberg, 1990: 105–7; 112–4). This might suggest the need to reconsider the proportions of fiction to nonfiction in an Australian corpus (in Brown/LOB the ratio is 1:3, whereas in the Nijmegen corpus described by Oostdijk (1988), it is 2:3). However the average Australian consumes quantities of nonfiction through periodicals (categories D – H) which were not covered in the Australia Council research, and their importance as part of the nation's reading profile is not to be underestimated.

One further point affecting the reception index of books in any field is competition from imported publications. By restricting the corpus to writing by Australians published in Australia, we turned our backs on the fact that what Australians read is often authored out of Australia, and published in Britain or North America. According to the Australia Council research, 70% of books bought, and 83% of those borrowed from libraries were from overseas (Guldberg 1990: 61, 65). Imported material might also be thought of as contributing to the melting pot of common usage, in Australia and in other English-speaking countries.

By making the Australian corpus homogeneously Australian, we know what we have: a data base which is generically matched with Brown and LOB, and can reasonably be compared with them. But future Australian corpora could well make more room for the many non-Australian contributors to our variety of English. Their inclusion would enhance the corpus in representing the heterogeneity of the Australian speech community as well as the local publishing scene – even though it would at the same time reduce the validity of comparisons with corpora from other English-speaking societies.

A single corpus can never satisfy all kinds of linguistic inquiry. But if carefully and systematically compiled, it is a reliable testing ground for some of the questions, and we at least know what kind of data our answers are based on.

Notes

1. The majority of the words and compounds which seem to have entered Australian English since 1981 are derived from North America. See the *Macquarie Dictionary of New Words* (1990) ed. S. Butler.

2. Comparisons with the frequencies of tag combinations in LOB (Johansson and Hofland, 1989) are just one such possibility.
3. Questions relating to the difference in the sampling year for the Australian corpus have been discussed elsewhere (see Peters, 1987; and Collins and Peters, 1988). As this paper shows, it would have been even more difficult to obtain the necessary samples in corresponding areas of publishing if we had decided to make our sampling year prior to 1986.
4. See the notes pp. 16-20 in the *Manual of Information to Accompany the Lancaster-Oslo/Bergen Corpus of British English.*
5. See *ICAME News* 10 for a bibliography up to 1985, and Johansson and Stenström (1991) for publications since then.
6. The totals have been compiled from information on the selections made for LOB and Brown, given in their respective *Manuals*.

References

Butler, S. ed. 1990. *The Macquarie Dictionary of New Words*. Sydney: Macquarie Library.

Collins, P. and Peters, P. 1988. The Australian corpus project, in *Corpus Linguistics Hard and Soft*, ed. M. Kytö, O. Ihalainen and M. Rissanen. Amsterdam: Rodopi.

Daniel, H. 1989. *The Good Reading Guide*. Melbourne: McPhee Gribble.

Francis, W.N. 1982. Problems of assembling and computerising large corpora, in *Computer Corpora in English Language Research*, ed. S. Johansson. Bergen: Norwegian Computing Centre for the Humanities.

Gelder, K. and Salzmann, P. 1989. *The New Diversity: Australian Fiction 1970–1988*. Melbourne: McPhee Gribble.

Guldberg, H. 1990. *Books – Who Reads them?* Sydney: Australia Council.

Johansson, S. and Hofland K. 1989. *Frequency Analysis of English Vocabulary and Grammar* 2v. Oxford: Clarendon.

Johansson, S. and Stenström, A. (eds.). 1991. *English Computer Corpora: Selected Papers and Bibliography*. Berlin: Mouton de Gruyter.

Knight, S. 1990. *Crimes for a Summer Christmas*. Sydney: Allen and Unwin.

Krogvig, I. and Johansson, S. 1984. *Shall* and *will* in British and American English: A frequency study, *Studia Linguistica* 35, 70–87.

Leitner, G. 1991. The Kolhapur Corpus of Indian English – intravarietal description and/or intervarietal comparison, in Johansson and Stenström (1991).

Manual of Information to Accompany the Lancaster – Oslo/Bergen Corpus of British English. Department of English, University of Oslo, 1978.

Manual of Information to Accompany a Standard Corpus of Present-day Edited American English. Revised and amplified. Department of Linguistics, Brown University, 1979.

Oostdijk, N. 1988. A corpus for studying linguistic variation, *ICAME Journal* 12, 7–14.

Peters, P. 1987. Towards a corpus of Australian English, *ICAME Journal* 11, 27–38.

Shastri, S. 1985. Word frequencies in Indian English: A preliminary report, *ICAME News* 9, 38–44.

Shastri, S. 1988. The Kolhapur Corpus of Indian English, and work done on its basis so far, *ICAME Journal* 12, 15–26.

Taking a Parsed Corpus to the Cleaners: The EPOW Corpus

Tim F. O'Donoghue
University of Leeds

1. Background: The POW Corpus

The POW Corpus is a parsed corpus of childrens' spoken English. It was originally collected by Robin Fawcett and Michael Perkins between 1978–84 for a child language development project to study the use of various syntactico-semantic constructs in children between the ages of six and twelve. A sample of approximately 120 children in this age range from the Pontypridd area in South Wales was selected, and divided into four cohorts of 30, each within three months of the ages 6, 8, 10, and 12. These cohorts were subdivided by sex and by socio-economic class, using details of the 'highest' occupation of the child's parents and the parents' educational level.

The children were selected in order to minimise any Welsh or other second language influence. The above subdivision resulted in small homogeneous cells of three children. Recordings were made of a play session for each cell, and of an individual interview with the same 'friendly' adult for each child, in which the child's favourite games or television programmes were discussed. The first 10 minutes of each play session commencing at a point where normal peer group interaction began (i.e. when the microphone was ignored) were transcribed by 15 trained transcribers. Likewise for the interviews. Transcription conventions were adopted from those used in the Survey of Modern English Usage at

University College London, and a similar project at Bristol. Intonation contours were added by a phonetician to produce a hard copy version, and the resulting transcripts published in four volumes (Fawcett et al. 1980). A short report on the project was also published (Fawcett 1980).

Of the approximately 100,000 words of speech transcribed, 61,000 were manually parsed giving 11,000 sentences. The parsing scheme used was Fawcett's development of Systemic Functional Grammar (SFG), full details of which can be found in Fawcett et al. 1981 and, more accessibly in Butler 1985: 94–101. A fuller description of the corpus itself can be found in Souter 1989b while a discussion of systemic functional corpora in general can be found in Souter 1990.

2. *Errors in the POW Corpus*

Unfortunately the parsed corpus contains a great number of errors and inconsistencies (as is to be expected in anything of this scale that is done by hand). This is despite the fact that a rigorous procedure for making the analysis was followed. Ten graduate analysts were trained by Fawcett; each analysed a text first on their own, then had it cross-checked by a partner analyst. In the first weeks all texts were checked by Fawcett. Soon uniform analyses were being achieved, but Fawcett continued holding regular 'problem sessions' with the analysts throughout. Finally all texts were checked a third time for uniformity by the senior analyst, who consulted Fawcett when difficulties arose. Most of the problems, however, concerned alternative analyses, and even if some persisted they would not show up as inconsistencies in the computer analysis. Factors contributing to the problems for computational analysis included:

- In the textual analyses, a distinction was made between upper and lower case letters for grammatical category labels (for example, *c* and *C* are distinct grammatical category labels in the textual analyses). However when the analyses were transferred to the computer, only upper case letters were used so a mapping had to be devised which ensured that the original mixed-case labels mapped onto distinct upper-case equivalents (for example, *C* remained as *C* while *c* was transformed to *CV*). Human error in typing these in led to some changes not being made – this gives rise to erroneously labelled parse trees in the machine readable version of the corpus.

- Other errors were introduced at the stage of entering the analyses in the computer. For the machine readable form, a novel numerical format was devised to encode the textual parse trees. Unlike a bracketed encoding of a parse tree, this format was designed to capture the discontinuity that occasionally occurs in sentence analysis. However, the complexity of this encoding led to bad encodings of the textual analyses, giving rise to many illegal structures in the machine readable version.

3. Motivation for editing the corpus

Though the machine readable version contains many errors and inconsistencies it is still a valuable parsed corpus resource whose parse trees are richly labelled with the syntactico-semantic categories of systemic syntax. The information content of the POW Corpus is high relative to other existing parsed corpora (for example, the LOB Treebank (Sampson 1987), which is parsed with respect to a theory-neutral model of syntax, yielding a very 'surfacy' structural analysis of a sentence). As such POW has been used by corpus linguists, most recently in the COMMUNAL project for grammar extraction (Atwell et al. 1988, Souter 1989a, Souter 1990). Although a context-free phrase structure grammar was not used in the analysis of the POW Corpus, it is possible to extract such a grammar simply by collecting the context-free phrase structure rules that would have to be applied for each tree in the corpus. For example, Atwell 1988 shows how this was done with the LOB Treebank. However, the grammars extracted from the POW Corpus contain many erroneous rules as a result of the errors in the corpus. It was the need for a clean corpus on which such extraction techniques could be performed that provided the motivation to edit the POW Corpus.

4. Detecting and editing errors

The structure of well-formed systemic functional parse trees obeys strict immediate dominance relationships. This results in layered parse trees in which there are alternating levels of different types of grammatical categories: nodes labelled with formal syntactic categories (*units*) such as clause (*CL*) and nominal group (*NGP*) immediately dominating a layer of nodes labelled with functional labels (*elements*) such as main verb (*M*) and head (*H*). For example, consider the parse tree in Figure 1. It is layered from the root down with elements, units, elements,..., units, elements and finally words. Each type of unit has a set of potential elements (functionally labelled nodes that it can immediately dominate), for example:

- A clause can dominate 49 different elements including subject (*S*), main verb operator (*OM*) and complement (*C*).
- A nominal group can dominate 28 different elements including deictic determiner (*DD*), quantifying determiner (*DQ*),pronominal head (*HP*) and head (*H*).
- Other units such as preposition groups, quantity-quality groups and genitive clusters similarly have their own sets of potential elements. Definitive lists of these potential elements can be found in the POW Handbook.

Certain elements are terminal grammatical categories and as such immediately dominate ('are directly expounded by') words, e.g. formula (*F*) is expounded by "YEAH". Other elements are non-terminal grammatical categories and immediately dominate ('are filled by') further units, the sentence (*Z*) is filled by two clauses.

Thus a number of immediate dominance relationships can be defined for systemic functional parse trees; for example: words must be immediately dominated by a lexical element, units must be immediately dominated by a higher element of structure, main verbs must be immediately dominated by a clause and so on. Such immediate dominance relationships must be obeyed by any valid systemic functional parse tree and so can be used as the basis for a procedure to check systemic functional parse trees. These valid relationships, obtained from data in the POW Handbook, were encoded into a procedure to automatically check each parse tree in the corpus and highlight any areas in parse trees which violated those valid relationships. This procedure detected errors in 900 (8.2%) of the 11026 parse trees that make up the POW Corpus. These 900 parse trees were edited manually and then checked by a trained linguist who was familiar with the corpus (Clive Souter at the School of Computer Studies in Leeds University) to ensure that no errors had been introduced during the editing process (either through incorrectly editing an existing error or accidentally introduced a new one). This editing process is described in detail in O'Donoghue 1991.

After editing, the corpus underwent a spelling check in which the words labelling the leaves in the trees were checked by the UNIX spell checker. 696 words were deemed to have been spelled incorrectly. However, 435 of these words were in fact legal words; they were merely proper nouns or words not covered by the vocabulary of the spelling checker. The remaining 261 truly incorrectly spelled words were then corrected.

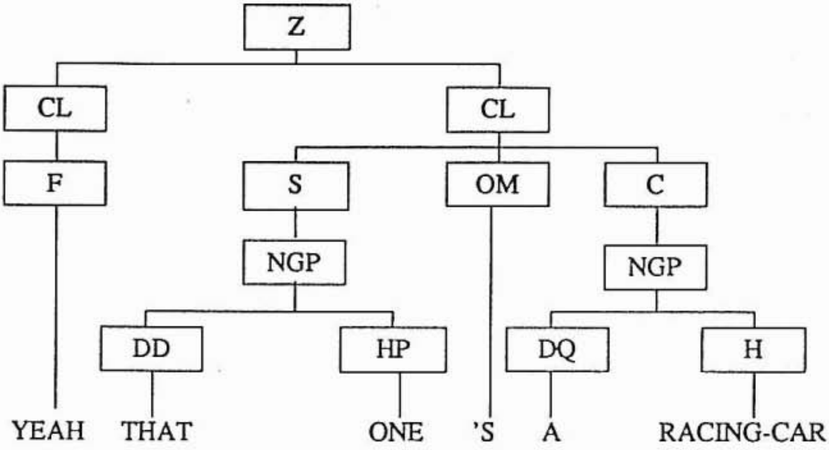


Figure 1: A typical POW parse tree

5. The result: A 'cleaner' corpus

The result of editing the corpus can be gauged by comparing the number of observed context-free phrase structure rules in the corpus before and after editing. Even though a context-free phrase structure grammar was not used in the analysis of the corpus, such a model can provide an indication of the range of syntactic structures contained in the corpus. The number of unique observed rules in POW (before editing) is 4340 compared with 3835 in EPOW (after editing). If we assume that the EPOW rule set is 'correct', then any rules which appear in the POW set but not in the EPOW set can be considered to be erroneous. Under this assumption, 593 (13.7%) unique context-free rules observed in the POW Corpus were erroneous. Of these 593 erroneous rule types, one rule type occurred 76 times while 97/593 (16.4%) of the erroneous rule types occurred more than once. The structures giving rise to these erroneous rules have been corrected in the EPOW Corpus thus providing a parsed corpus resource in which corpus linguists can have more faith – faith that a rare rule is truly a rare rule and not some artifact introduced in the machine readable form.

6. Availability

The EPOW Corpus is currently being distributed by the author; however, distribution may be undertaken by ICAME sometime in the future. The parse trees in the EPOW Corpus come in three different forms:

- Encoded in the original numerical format (as described in O'Donoghue 1991 and illustrated in Souter 1989b and Souter 1990).
- Encoded as bracketed lists (trees with discontinuities have had their leaves re-ordered for the sake of computational tractability).
- Encoded as predicate-argument structures (here also trees with discontinuities have had their leaves re-ordered).

The parsed corpus is also distributed with a raw version of the corpus (i.e. sentences without parse trees) and a number of syntactic models extracted from the parse corpus (including context-free phrase structure rules and a lexicon, both with frequency data). PostScript versions of this paper, a paper describing the editing process in more detail (O'Donoghue 1991) and the original handbook for the POW Corpus (Souter 1989b) are also included in the distribution. Copies of the corpus and its associated files and documentation can be obtained by sending either a tape suitable for UNIX Tar or a 1/4in data cartridge for Sun Tar to the author. The conditions of distribution are the same as for the POW Corpus (see Souter 1989b for details).

References

Atwell, Eric Steven. 1988. "Transforming a Parsed Corpus into a Corpus Parser". In Kytö et al. (1988), 61–70.

Atwell, Eric Steven, D. Clive Souter, and Tim F. O'Donoghue. 1988. *Prototype Parser 1*. COMMUNAL Report 17, School of Computer Studies, University of Leeds.

Butler, Christopher S. 1985. *Systemic Linguistics: Theory and Applications*. London: Batsford.

Fawcett, Robin P. 1980. "Language Development in Children 6–12: Interim Report". *Linguistics*, 18:953–958.

Fawcett, Robin P. 1981. *Some Proposals for Systemic Syntax*. Department of Behavioural and Communication Studies, Polytechnic of Wales, 1981. Previously published in the *Journal of the Midlands Association for Linguistic Studies (MALS)*, Volumes 1–2, 2–1, 2–2 (1974–76).

Fawcett, Robin P. and Michael R. Perkins. 1980. *Child Language Transcripts 6–12 (with a Preface, in 4 Volumes)*. Department of Behavioural and Communication Studies, Polytechnic of Wales, 1980.

Garside, Roger, Geoffrey Sampson, and Geoffrey Leech (eds.) 1987. *The Computational Analysis of English: A Corpus Based Approach*. London: Longman.

Kytö, Merja, Ossi Ihalainen, and Matti Rissanen (eds.) 1988.

Corpus Linguistics, Hard and Soft: Proceedings of the 8th International Conference on English Language Research on Computerized Corpora. Amsterdam: Rodopi.

Meijs, Willem (ed.). 1990. *Theory and Practice in Corpus Linguistics.* Amsterdam: Rodopi.

O'Donoghue, Tim F. 1991. "EPOW: The Edited Polytechnic of Wales Corpus". To appear in *Proceedings of the 5th International Conference on Symbolic and Logical Computing (ICEBOL5)*, Dakota State University, 1991.

Sampson, Geoffrey. 1987. "The Grammatical Database". In Garside et al. (1987), 82–96.

Souter, Clive. 1989a. "The COMMUNAL Project: Extracting a Grammar from the Polytechnic of Wales Corpus". *ICAME Journal*, 13:20–27.

Souter, Clive. 1989b. *A Short Handbook to the Polytechnic of Wales Corpus*. Bergen: Norwegian Computing Centre for the Humanities.

Souter, Clive. 1990. "Systemic Functional Grammars and Corpora". In Meijs (1990), 179–211. Also available as *School of Computer Studies Research Report* 89.12, University of Leeds.

Tagging Brown with the LOB Tagging Suite

Nancy Belmore
Concordia University

1. Purpose

My purpose in using these corpora is somewhat different from that of most other users. The aim is to use them as a research tool in arriving at a word classification system which has a sufficient number of classes and sub-classes to be adequate for natural language understanding systems. The number of such classes is much larger than would normally be considered adequate for grammatical description and all of the classes should be what Sager (1981, p. 9) has called 'informationally relevant'. Of course my hope is that using the corpora for this purpose will also contribute to the 'better general linguistic descriptons' to which Svartvik (1987, p. 36) alluded in his paper 'Taking a New Look at Word Class Tags'.

There are numerous ways in which tagged corpora can be used to achieve this aim. In an initial pilot study at the University of Amsterdam (Belmore, 1987), I used the QUERY program (Meijs, 1982; Van der Steen, 1982 and 1984) to determine the usefulness of pattern extraction from tagged corpora in order to get precise information on the exact circumstances in which problems in defining particular word classes arise and what would be required to resolve them. In another, at the University of Lancaster (Belmore, 1988a), I used the original LOB tagging suite, CLAWS1 (Garside, 1987), to tag a subset of sentences exemplifying some of the major problems the QUERY output had revealed. CLAWS1 was used to tag the LOB Corpus available from the International Computer Archive of Modern English in Bergen, Norway (hereafter ICAME).

The two studies demonstrated the value of locating authentic examples of problematic tagging decisions through pattern extraction from tagged corpora. They also showed the value of finding out how different tagging systems tag the same set of problematic examples, or even examples which do not appear to present a problem. In the latter case, unanticipated differences in tagging decisions may reveal unexpected insufficiencies in word class definitions.

I therefore decided to undertake tagging the entire Brown Corpus with CLAWS1. This will make it possible to compare objectively the differences between the Brown Corpus as tagged by CLAWS1 and the Brown Corpus as originally tagged at Brown University. My aim is to achieve this, insofar as possible, from my own desktop, using a MacIntosh II. The advantages are numerous and are described in some detail in Belmore (1988b).

2. *Procedures*

I had already downloaded, edited and reformatted a sample file from the tagged Brown Corpus (Belmore, 1988b), chosen because it was the smallest file in the corpus and yet large enough – about 2500 words – to provide adequate test data. I reduced the size of the file by 75% by the simple expedient of removing trailing spaces from each field of each record. Figure 1 shows an extract from this file.

Figure 1. Extract from a file from the tagged Brown Corpus which has been edited and reformatted.

Oslo	np-hl	A04001001E1
the	at	A04001002E1
most	ql	A04001003E1
positive	jj	A04001004E1
element	nn	A04001005E1

to	to	A04001006E1
emerge	vb	A04001007E1
from	in	A04001008E1
the	at	A04001009E1
Oslo	np	A04001010E1
meeting	nn	A04001011E1
of	in	A04002001E1
North	jj-tl	A04002002E1
Atlantic	np-tl	A04002003E1
Treaty	nn-tl	A04002004E1
Organization	nn-tl	A04002005E1
Foreign	jj-tl	A04002006E1
Ministers	nns-tl	A04002007E1
has	hvz	A04002008E1
been	ben	A04002009E1
the	at	A04003001E1
freer	jjr	A04003002E1
,	,	A04003003E1
franker	jjr	A04003004E1
,	,	A04003005E1
and	cc	A04003006E1
wider	jjr	A04003007E1
discussions	nns	A04003008E1
,	,	A04003009E1
animated	vbn	A04003010E1
by	in	A04003011E1
much	ql	A04003012E1
better	jjr	A04003013E1
mutual	jj	A04003014E1
understanding	nn	A04004001E1
than	cs	A04004002E1
in	in	A04004003E1
past	jj	A04004004E1
meetings	nns	A04004005E1
.	.	A04004006E1

I later downloaded the untagged counterpart of this file from the typographically enhanced version of the corpus available from

ICAME called Bergen I. In this version, all upper-case letters which were lower-case in the texts from which the original Brown Corpus was keyed in have been replaced by lower-case. This is a fortuitous enhancement because CLAWS1 expects upper- and lower-case letters. Figure 2 shows an extract from this file (hereafter BrnB) as it appeared when first downloaded.

Figure 2. Extract from a file from the untagged Brown Corpus when first downloaded. The file is called BrnB.

```
A04 0010 OSLO The most positive element to emerge from the Oslo meeting
A04 0020 of North Atlantic Treaty Organization Foreign Ministers has been
A04 0030 the freer, franker, and wider discussions, animated by much better mutual
A04 0040 understanding than in past meetings.  This has been a working
A04 0050 session of an organization that, by its very nature, can only proceed
A04 0060 along its route step by step and without dramatic changes. In Oslo,
A04 0070 the ministers have met in a climate of candor, and made a genuine
A04 0080 attempt to get information and understanding one anot-her's problems.
A04 0090 This atmosphere of understanding has been particularly noticeable
A04 0100 where relations are concerned between the "colonialist" powers
A04 0110 and those who have never, or not for a long time, had such problems.
```

2.1 Editing

The major task was to edit BrnB so that it would conform to the expectations of CLAWS1. This requires knowledge of the relations between the ICAME coding conventions for Bergen I, the conventions for coding the original Brown Corpus and the conventions used in preparing the LOB Corpus for input to CLAWS1. It also requires knowledge of the extensive manual editing of the LOB Corpus which occurred after running the PREEDIT program, the first program in CLAWS1. Several different types of editing were necessary:

1. Restoration of information no longer explicitly indicated.
   In Bergen I, e.g., a number of compound symbols in the original corpus have been replaced by a single character, sometimes suppressing information, like the difference between a begin-quote and an end-quote, which CLAWS1 expects. In such cases, I had to restore the original Brown compound symbols.

2. Insertion of meta-symbols which CLAWS1 expects, e.g., a sentence initial marker and a paragraph marker.

3. Prevention of clashes.
   In Bergen I, the compound symbol to indicate a period marking an abbreviation has been replaced by an ampersand, which CLAWS1 will interpret as representing itself and tag as a coordinating conjunction. Similarly, a '~' indicates an acronym while to CLAWS1 it indicates an included sentence.

   Some potential clashes were more subtle. In both Brown and LOB a hyphen within a word is properly construed as part of that word. This is the only use of a hyphen in Brown. In LOB, however, a hyphen may also occur surrounded by spaces. Under appropriate circumstances it will be tagged as a preposition. Johansson (1986, p. 119) gives as an example a sequence like 4 1/2 – 6. A search for all hyphens in BrnB showed that none had occurred surrounded by spaces. However, as the examples in Figure 3 show, the context in which some of them occurred, not within a word, but at the end, indicated that they were not meant to be interpreted as part of the word. If so, should I insert a space before each one and then see how LOB would tag them? Fortunately, I decided to look at the corresponding Brown tagged examples where I discovered that all of the hyphens in question had originally been coded as dashes. I learned in this way that, to avoid a clash, the hyphens in Brown must be recoded as dashes when they represent a mark of punctuation, but not when they represent a preposition.

4. Reformatting.
   LOB expects headings and paragraphs to begin on a separate line, but in Bergen I they do not necessarily do so. To reduce the size of the file and to simplify searches, I also decided to replace one or more single spaces by a single space.

Figure 3. All lines with one or more hyphens in the file BrnB. The original file has already been considerably edited at this point.

A03 0750	*WASHINGTON, FEB& 9* – President Kennedy today proposed a mammoth
A03 0890	per cent- 3 per cent on each worker and employer- on the first $4,800
A03 0910	$5,000 a year and the payroll tax to 6.5 per cent- 3.25 per cent each.
A03 1090	a hospital. A patient could receive up to 300 days paid-for nursing
A03 1110	who use none or only part of the hospital-care credit. *3*. Hospital
A03 1190	staggered by the drain on their savings- or those of their children-
A03 1370	to $1,500 a year for one-fourth of the first year students. The
A03 1380	schools could use the money to pay 4-year scholarships, based on need,
A03 1620	million dollars in the 1961–62 budget for direct government research
A03 1640	elements in a sound health program- people, knowledge,
A03 1670	lines. Legislators who last year opposed placing aged-care under
A03 1770	*WASHINGTON, FEB& 9* – Acting hastily under White House pressure,

5. Performing automatically, *before* running the PREEDIT program, as many as possible of the manual editing tasks which occurred after running the PREEDIT program when the tagged LOB corpus was prepared.

   In the original corpus, e.g., what was called an item-number/letter and a sentence immediately following it were coded as a single sentence. Then, during the manual editing following PREEDIT, a period was inserted, if it happened to be missing, after the item-number/letter and a sentence initial marker was inserted immediately before the next printable character. In other words, what had orignally been coded as one sentence was now coded as two. Since BrnB marks such constructions, it was possible to do this editing at the outset.

   There were other similar instances. PREEDIT looks for words with more than one upper-case letter and replaces them by lower-case letters. Then a manual editor must restore those capitals which shouldn't have been removed. The instances in which this step is required can be minimized by finding all such words in the input data and making the required changes at that point. In some cases, correct tagging requires changing all the capitals to lower-case while in others the first character should remain upper-case. Figure 4 shows some of the words in this category which occurred in BrnB.

Figure 4. An extract from a search file. The file contains all lines from BrnB with words containing one or more capitals.

```
A03 0770 American workers would be raised to pay the hospital and some other
A03 0780 medical bills of 14.2 million Americans over 65 who are covered by
A03 0860 *<COST UP TO $37 A YEAR*>
A03 1660 *<REACTION AS EXPECTED*>
A04 0020 of North Atlantic Treaty Organization Foreign Ministers has been
A04 0120 The nightmare of a clash between those in trouble in Africa, exacerbated
```

A04 0161 *<EXPLOSION AVOIDED*>
A04 0180 should critics of its Angola policy prove harsh, there has been
A04 0220 in the ~UN General Assembly as to ~NATO members' votes,
A04 0240 in the future such topics as Angola will be discussed in advance.

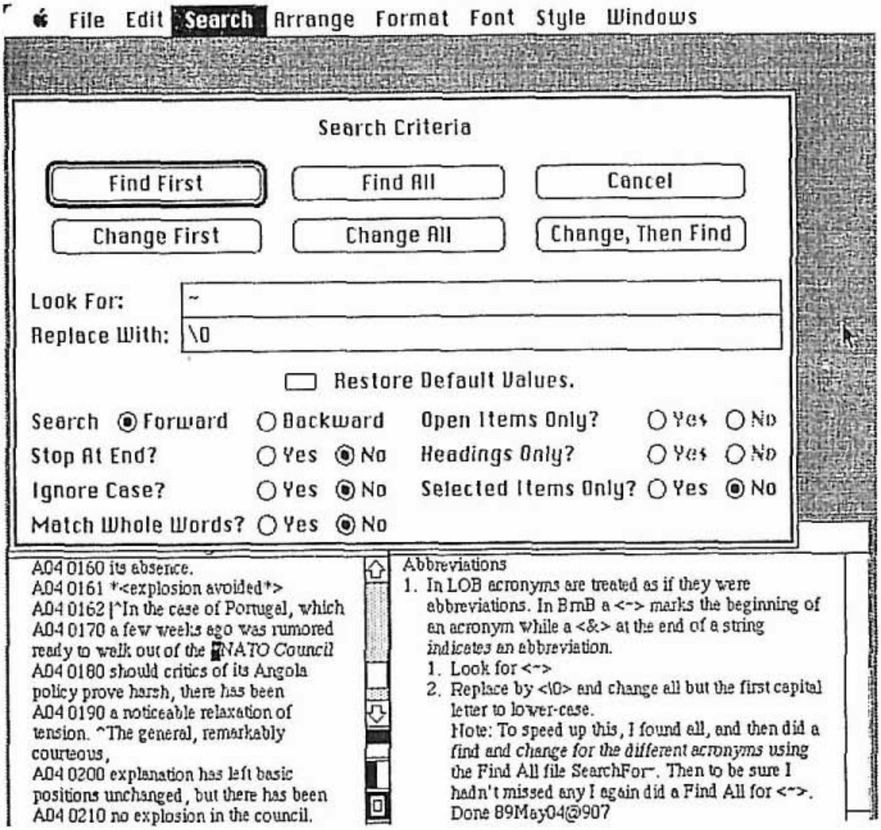
Doing the editing required a word processor with sophisticated search and replace capabilities. MindWrite is just such a word processor. For each step in the editing, I noted in an editing document the basis for the decision to perform that particular step, whether *The Tagged LOB Corpus Users' Manual* (Johansson, 1986), the manual of information for Brown (Francis and Kučera, 1979), the one for LOB (Johansson, Leech and Goodluck, 1978), the *LOB Manual PRE-EDIT Handbook* (Atwell, 1981) or documentation from ICAME. I also recorded the search criteria, the replacement string and the date any searches or replacements were actually effected. In addition, I kept a document with general comments and questions and, if the editing followed a test run of the PREEDIT program, yet another document with comments on the results of that particular run.

MindWrite allows opening more than one document at the same time so that I could fill in the search criteria in the editing document, and 'paste' them into the 'Look For' box in the MindWrite Search Criteria 'window' in the current version of BrnB. I then checked to see if the required string would be found, revising the criteria if necessary – it is surprisingly easy to state the criteria inadequately or incompletely, e.g., to forget to state them so that the 8-digit reference number will be skipped. The next step was to check the accuracy of the replacement string and, if satisfied, to replace all the strings meeting the search criteria with that string. Figure 5 (p. 72) is a screen snapshot from a typical editing session.

Before doing a 'Replace With', it was often prudent and sometimes essential to do a 'Find All'. A 'Find All' collects all instances found in a temporary document which is displayed in a Search window. All searches doing a particular session are consecutively numbered, and each such search can be saved as a separate document. I have made a practice of saving the results of searches whenever I thought it could be important, or just interesting, to have a record of the strings to be modified. For the same reason, after doing a 'Replace All', I would often do a 'Find All' and then save that document so that I could have a record of what had been changed and of what the effect of the change was. Figure 6 shows all lines with begin quotes in BrnB. Before this search began, all quotes in strings which implied they were end quotes were temporarily replaced by the string <**æ>. This was an intermediate step which simplified finding the begin quotes: they were any quotes remaining in the document after all end quotes had been found and temporarily represented as <**æ>. Figure 7 shows an extract from a search file containing all lines with one or more begin or end quotes. Quotes in this search file have been replaced by curly quotes for readability whereas in BrnB the LOB compound symbols occur instead.

(cont. on p. 73)

Figure 5. A screen 'snapshot' from a typical editing session. The only difference is that the borders of the document 'windows' were chosen to be temporarily smaller so that the MindWrite search 'window' could be shown together with them. The pair of 'scrollable' document 'windows' can fill the screen and the user can make the search 'window' appear or disappear 'in front' of them. The document in the lower left is the version of BrnB being edited; the one on the right, the editing document.



Look For: ~

Replace With: \0

□ Restore Default Values.

Search ● Forward ○ Backward

Stop At End? ○ Yes ● No

Ignore Case? ○ Yes ● No

Match Whole Words? ○ Yes ● No

Open Items Only? ○ Yes ○ No

Headings Only? ○ Yes ○ No

Selected Items Only? ○ Yes ● No

A04 0160 its absence.
A04 0161 *<explosion avoided*>
A04 0162 |^In the case of Portugal, which
A04 0170 a few weeks ago was rumored
ready to walk out of the NATO *Council*
A04 0180 should critics of its Angola
policy prove harsh, there has been
A04 0190 a noticeable relaxation of
tension. ^The general, remarkably
courteous,
A04 0200 explanation has left basic
positions unchanged, but there has been
A04 0210 no explosion in the council.

Abbreviations

1. In LOB acronyms are treated as if they were abbreviations. In BrnB a <~> marks the beginning of an acronym while a <&> at the end of a string *indicates an abbreviation.*
   1. Look for <~>
   2. Replace by <\0> and change all but the first capital letter to lower-case.
      Note: To speed up this, I found all, and then did a *find and change for the different acronyms using* the Find All file SearchFor~. Then to be sure I hadn't missed any I again did a Find All for <~>.
      Done 89May04@907

*Figure 6. An extract from a search file containing all lines with begin quotes in BrnB. Before this search began, all quotes in strings which implied they were end quotes were temporarily replaced by the string <**æ>.*

A03 0541 |^Karns said it was a "wrongful act**æ
A03 0550 for Wexler to take statements "privately and outside of the grand
A03 0560 jury room**æ. ^He said this constituted a "very serious misuse**æ
A03 0571 |^"Actually, the abuse of the
A03 0640 "complementary**æ miscount of the vote, in which votes would be taken
A03 1100 home care under a "unit formula**æ allowing more of such care for those

Figure 7. An extract from a search file containing all lines with one or more begin or end quotes. Quotes in this search file have been replaced by curly quotes for readability whereas in BrnB the LOB compound symbols occur instead.

A03 0541 |^Karns said it was a 'wrongful act'
A03 0550 for Wexler to take statements 'privately and outside of the grand
A03 0560 jury room'. ^He said this constituted a 'very serious misuse'
A03 0571 |^'Actually, the abuse of the
A03 0600 function of this court', said Karns, who is a City judge in East
A03 0640 'complementary' miscount of the vote, in which votes would be taken
A03 1100 home care under a 'unit formula' allowing more of such care for those

Since a 'Replace All' can be a drastic step, I would usually – unless very confident – save the document before doing a 'Replace All'. Then if the result turned out to be what Charles Shulz's Snoopy called 'a rude awakening', I could do a 'Revert to Saved' and restore the document to its status immediately preceding the miscalculated 'Replace All'.

An essential part of the editing is to go through the steps in the right order. Thus before replacing more than one space by a single space, it is essential to identify the beginning of a paragraph and insert the LOB paragraph marker. This is because in the original file from Bergen multiple spaces are used to mark paragraphs. A sentence closer followed by three spaces signals a paragraph break within a record; when a paragraph begins on a new record, there are three spaces in positions 10–12.

What I have done so far is what I call pre-programming, i.e., based on documentation and data from various sources, I have specified each editing step and when it should be carried out and then used used MindWrite to do the editing interactively. I call this pre-programming because, although it would be impossible to do what I have done manually, and I am reasonably certain that the editing steps so far are accurate and in the right order, they must ultimately be incorporated into a program which can be run over the entire set of files without my intervention.

I have checked the adequacy of the editing by running PREEDIT with successive versions of the edited sample as input. Most of the required editing was completed before the first test. A few more steps were completed before the second. The only change required before running the third test was to replace hyphens by dashes, as described above.

Since this was a sample in which, e.g., only one colon occurred and no semi-colons, certain editing steps that will ultimately be required could be postponed. Figure 8 shows an extract from BrnBafter the editing had been completed. It should be compared with the excerpt from the original unedited file shown in Figure 2. Figure 9 shows an extract from the third output of PREEDIT.

Figure 8. An extract from BrnB after the editing had been completed. It should be compared with the extract from the original unedited file shown in Figure 2.

A04 0010 |^Oslo. ^The most positive element to emerge from the Oslo meeting

```
A04 0020 of North Atlantic Treaty Organization Foreign Ministers
         has been
A04 0030 the freer, franker, and wider discussions, animated by
         much better mutual
A04 0040 understanding than in past meetings.
A04 0041 |^This has been a working
A04 0050 session of an organization that, by its very nature, can
         only proceed
A04 0060 along its route step by step and without dramatic
         changes.
A04 0070 ^In Oslo, the ministers have met in a climate of candor,
         and made a genuine
A04 0080 attempt to get information and understanding one anot-
         her's problems.
A04 0090 ^This atmosphere of understanding has been particularly
         noticeable
A04 0100 where relations are concerned between the *"colonia-
         list**" powers
A04 0110 and those who have never, or not for a long time, had
         such problems.
```

Figure 9. An extract from the third output of PREEDIT. Three progressively more adequate versions of BrnB were input to PREEDIT.

```
A04 0010020  oslo
A04 0010021  .                                   .
A04 0010022  ---------------------------------------------
A04 0010030  the
A04 0010040  most
A04 0010050  positive
A04 0010060  element
A04 0010070  to
A04 0010080  emerge
A04 0010090  from
A04 0010100  the
A04 0010110  Oslo
A04 0010120  meeting
```

```
A04 0020020  of
A04 0020030  North
A04 0020040  Atlantic
A04 0020050  Treaty
A04 0020060  Organization
A04 0020070  Foreign
A04 0020080  Ministers
A04 0020090  has
A04 0020100  been
A04 0030020  the
A04 0030030  freer
A04 0030031  ,                        ,
A04 0030040  franker
A04 0030041  ,                        ,
A04 0030050  and
A04 0030060  wider
A04 0030070  discussions
A04 0030071  ,                        ,
A04 0030080  animated
A04 0030090  by
A04 0030100  much
A04 0030110  better
A04 0030120  mutual
A04 0040020  understanding
A04 0040030  than
A04 0040040  in
A04 0040050  past
A04 0040060  meetings
A04 0040061  .                        .
A04 0041012  ------------------------------
```

2.2 Tagging

While I edited the test data on the MacIntosh II, a colleague at our Computer Centre, Anne G. Barkman, worked on making the necessary revisions in CLAWS1 so that it would run on a VAX (the original programs were run on an ICL). Only minor changes to PREEDIT were required. The succeeding programs also required changes, mostly because of differences in the EBCDIC and ASCII sorting sequence.[1]

The next-to-last program in CLAWS1, CHAINPROBS, assigns one or more tags, along with the probability that it is the correct tag, to each word. When run over the edited test data, this program reported that of the 2514 input words, 1709 had been unambiguously tagged by earlier programs in the tagging suite while 805 were still ambiguous. It resolved the ambiguity of a further 529, assigning more than one tag, together with a probability, to only 276 words. Figure 10 shows an extract from the output of this program. It corresponds to the sample output from the original tagged Brown Corpus shown in Figure 1 so that the original Brown tag for each word can be compared with the tagging from CHAINPROBS.

Figure 10. An extract from the output of CHAINPROBS. Compare the Brown tag for each word in Figure 1 with these tag(s).

```
A04 0010012 ------------------------------------------------------
A04 0010020 oslo             54 NN
A04 0010021 .                01 .
A04 0010022 ------------------------------------------------------
A04 0010030 the              02 ATI
A04 0010040 most             02 [QL]/ 85 AP/ 15 RBT@/  0
A04 0010050 positive         02 JJ
A04 0010060 element          02 NN
A04 0010070 to               02 [TO]/ 94 IN/  6
A04 0010080 emerge           54 [VB]/ 94 NN/  6
A04 0010090 from             02 IN
A04 0010100 the              02 ATI
A04 0010110 Oslo             52 NP
A04 0010120 meeting          02 [VBG]/ 62 NN/ 38
A04 0020020 of               02 IN
A04 0020030 North            52 NP
A04 0020040 Atlantic         02 NP
A04 0020050 Treaty           52 NP
A04 0020060 Organization     47 NNP
A04 0020070 Foreign          52 NP
```

```
A04 0020080 Ministers          48 NPTS
A04 0020090 has                02 HVZ
A04 0020100 been               02 BEN
A04 0030020 the                02 ATI
A04 0030030 freer              02 JJR
A04 0030031 ,                  01 ,
A04 0030040 franker            54 [NN]/ 70 JJR/ 30
A04 0030041 ,                  01 ,
A04 0030050 and                02 CC
A04 0030060 wider              02 JJR
A04 0030070 discussions        55 NNS
A04 0030071 ,                  01 ,
A04 0030080 animated           99 VBN
A04 0030090 by                 99 IN
A04 0030100 much               02 [AP]/ 66 RB/ 34
A04 0030110 better             02 [JJR]/ 56 RBR/ 37 VB%/ 6 NN%/ 1
A04 0030120 mutual             54 JJ
A04 0040020 understanding      02 [NN]/ 57 VBG/ 22 JJ@/ 21
A04 0040030 than               02 CS/ 58 [IN]/ 42
A04 0040040 in                 02 IN
A04 0040050 past               02 [NN]/ 79 IN/ 17 RI/  4
A04 0040060 meetings           54 NNS
A04 0040061 .                  01 .
A04 0041012 -------------------------------------------------
```

The final program in CLAWS1 assigns a single tag to each word. If no manual editing of the output of CHAINPROBS occurs, this tag is the tag CHAINPROBS designated the most likely. When the LOB Corpus was tagged, however, there was manual post-editing following CHAINPROBS whenever no one alternate had been assigned an acceptably high probability (Johansson, 1986, p. 21). In such cases, a tag other than the one CHAINPROBS had designated the most likely was sometimes selected, and it is possible to retrieve all such instances. This subset of tagged words could be another way of getting objective information about problematic tagging decisions. For my present purposes, however, I think the output of CHAINPROBS will be the most useful output.

3. Further Work

Before CLAWS1 can be run over the *entire* one-million word Brown Corpus, there is, of course, further work. Samples from other genres in the untagged corpus need to be edited with MindWrite until there is reasonable certainty that all necessary editing has been identified. After that, a program can be written to carry out the editing with almost no manual intervention.

3.1 Editing

As noted above, tagging the original LOB Corpus required extensive editing of the output from PREEDIT in order to produce a proper input for the succeeding programs. I have already done much of the editing which originally followed PREEDIT before running PREEDIT but I ignored for this particular test all the editing which is supposed to follow PREEDIT.

Some of this is fairly simple and obvious. Since PREEDIT removes a sentence initial capital (hereafter SIC), whenever an SIC has occurred on a word which is normally written with a word initial capital (hereafter WIC), such capital letters must be restored.

Far more problematic is the manual tagging of words with WICs. The purpose was to indicate several sub-classes of proper nouns. The directions that the editors followed were often complex and sometimes involved the application of fairly sophisticated linguistic criteria. In her description of the techniques the Lancaster group has since developed in order to eliminate such editing, Barbara Booth (1987, p. 101) noted that '...the major task of the human pre-editors was dealing with the problem of capitalization.'

The question which faces me is whether to attempt to repeat this phase of the editing or to follow the original Brown tagging procedure in this one respect. Of their decision to tag all words beginning with a capital *NP*, with the tag *TL* added to that tag whenever the word was part of a title, but with no further attempt at sub-classification, Francis and Kučera (1982, p. 13) have stated:

> The conventions of non-sentence-initial capitalization in English are complex and to a considerable degree variable, unlike most

> other aspects of the writing system. This has presented a problem in tagging, which has been disposed of, if not settled, by arbitary rules.

It seems to me that there was much merit in this decision and I have tentatively concluded that I will not attempt the elaborate manual sub-classification the LOB editors undertook. My reasoning is as follows:

Many of the rules the editors followed were based on punctuation conventions alone and, in fact, punctuation conventions sometimes took precedence over linguistic criteria. In other instances the criteria were semantic and subject to the inevitable problems the application of semantic criteria entail.

While I believe that users of the tagged LOB Corpus are pleased that this manual tagging was achieved – it is genuinely useful to know, e.g., when a proper noun functions like a common noun – my own purpose is to be able to retrieve words which present a problem in tagging. Where the Brown and LOB tagging systems normally agree, it is useful to pinpoint the exact contexts in which they result in different tagging decision. With proper nouns, however, it is known in advance that the decision will usually be different and that the differences arise from the application of manual tagging rules to arrive at sub-classes of the original Brown NP category.

As an alternative, it could be valuable to retrieve from the tagged LOB Corpus all instances of each sub-category of proper noun, linking them to the manual tagging rules which were apparently invoked, and revising them where this seems called for in order to achieve greater precision. To determine whether the revised rules can be applied consistently, at least two people could be asked to independently tag the same words in the untagged LOB Corpus. These results could then be compared to the results of the original manual tagging. This is a challenging area because it requires disentangling the role of variable punctuation conventions, semantic criteria and formal grammatical signals to arrive at rigorous word class definitions.

I have tentatively reached the same decision, on the basis of similar reasoning, with respect to manually inserting a sentence initial marker whenever a capitalized word follows a colon and to editing the original input data so as to indicate included sentences.

3.2 Electronic logbook

The Lancaster tagging group kept a logbook in which they recorded the decisions they had made during the editing following PREEDIT but the logbook no longer exists. Even if it were still available, like any document which is not computer-readable, it would be hard to use.

I have therefore, as illustrated earlier (see Figure 5), already begun an electronic logbook recording the exact steps in editing the Brown Corpus. I also plan to keep an electronic counterpart of the Lancaster logbook to record any changes that are made following the PREEDIT program, including documents showing the 'before' and 'after39 status of the corpus whenever this could be of value to me or other researchers. Van den Heuvel (1987, p. 247) has already noted the value of an electronic logbook in which manual interventions are systematically recorded.

3.3 4D Data Base

The most important task remaining is to set up a database in which to record the tagging decisions for each word in the corpus. I have described elsewhere (Belmore, 1988b) the advantages of the very powerful Fourth Dimension (4D) database management and programming language from ACIUS. A 4D database will facilitate answering two key questions: when is the tag which CLAWS1 designates the most likely tag different from the original Brown tag? When it is different, is the original Brown tag an alternative CLAWS1 tag?

4. Conclusion

Stig Johansson noted in *The Tagged LOB Corpus Users' Manual* (1986, p. 26):

> A particular problem has been that we have chosen to draw a borderline and assign a single tag for each occurrence of a

> word, though we know that gradience and fuzzy borderlines are characteristic of language.

He also observed (1986, p. 26)

> While an attempt has been made to find a classification which is linguistically justifiable, this has not always been possible. For one thing, this would have meant tackling grammatical problems which are still awaiting a solution.

To my knowledge, this is the first time an objective method has been developed and tested for determining the exact differences between two algorithms for grammatical analysis. My hope is that by making it possible to compare objectively the results of two different tagging procedures applied to the same corpus, we will have yet another tool to help us state more precisely where the fuzzy borderlines are and to make further progress in solving the many grammatical problems which are still awaiting a solution.

Acknowledgements

I would like to thank Geoffrey Leech and his colleagues from the Unit for Computer Research on the English Language (UCREL) at the University of Lancaster for their help in working with CLAWS1, most particularly Steve Elliott, who sent some crucial files by electronic mail and has always been available to answer questions. I would also like to thank Merja Kytö and Hannu Hartikka for answering questions about their experiences in running CLAWS1 on their VAX at the University of Helsinki. Very special thanks are due to Anne G. Barkman of Concordia University's Computer Centre, who converted the LOB and Brown files so that they would conform to VAX standards and made the necessary revisions to the CLAWS1 programs so that they would run on a VAX. The research described here has been partially supported by the Social Sciences and Humanities Research Council of Canada, to whom grateful acknowledgement is made.

Note

1. Just after Anne Barkman had completed the editing of the CLAWS1 programs, we learned that Tom Horton, who had

revised the programs to run on a VAX at the University of Edinburgh and whose notes she had used in making revisions for our installation, had made his revised programs available to researchers through UCREL at the University of Lancaster. In future runs, we will probably use that version. It is the version at the University of Helsinki and the one which researchers would now be given and we think it is important for all researchers to be using the same version, if possible. When we got CLAWS1 at Concordia, only the ICL version was available.

List of Software

MindWrite
DeltaPoint (formerly Access Technology, Inc.)
555C Heritage Harbor
Monterey, California 93940–2483

4th Dimension
ACIUS
Suite 495
20300 Stevens Creek Blvd.
Cupertino, California 95014

Canadian distributor of 4th Dimension:
Agoratech Canada, Inc.
Ste.–Thérèse, Québec, Canada
J7E 1W9

References

Atwell, Eric Steven. 1981. LOB Manual PRE-EDIT Handbook. Unpublished manuscript, Departments of Linguistics and Computer Studies, Lancaster Univ.

Belmore, Nancy F. 1987. 'A Pilot Study of the Application of Corpus Linguistics to the Specification of Word Classes for Language Understanding Systems'. In Willem Meijs (ed.), *Corpus Linguistics and Beyond*. Amsterdam: Rodopi, pp. 141–150.

Belmore, Nancy F. 1988a. 'The Use of Tagged Corpora in Defining Informationally Relevant Word Classes'. In Merja Kytö, Ossi Ihalainen and Matti Rissanen (eds.), *Corpus Linguistics, Hard and Soft*. Amsterdam: Rodopi, pp. 71–82.

Belmore, Nancy F. 1988b. 'Working with Brown and LOB on a Microcomputer', *Computer Corpora des Englischen*, 2 (2):1–14.

Booth, Barbara. 1987. 'Text Input and Pre-processing: Dealing with the Orthographic Form of Texts'. In Roger Garside, Geoffrey Leech and Geoffrey Sampson (eds.), *The Computational Analysis of English: A Corpus-based Approach*. London: Longman Group Ltd., pp. 97–109.

Francis, W. Nelson and Henry Kučera. 1982. *Frequency Analysis of English Usage: Vocabulary and Grammar*. New York: Houghton Mifflin Co.

Francis, W. Nelson and Henry Kučera. 1979. *Manual of Information to Accompany a Standard Sample of Present-day Edited American English, for Use with Digital Computers*. Original ed. 1964, revised 1971, revised and augmented 1979, Providence, R.I.: Department of Linguistics, Brown University.

Garside, Roger. 1987. 'The CLAWS Word-tagging System'. In Roger Garside, Geoffrey Leech and Geoffrey Sampson (eds.), *The Computational Analysis of English: A Corpus-based Approach*. London: Longman Group Ltd., pp. 30–41.

Heuvel, Theo van den. 1987. 'Interaction in Syntactic Corpus Analysis'. In Willem Meijs (ed.), *Corpus Linguistics and Beyond*. Amsterdam: Rodopi, pp. 235–252.

Johansson, Stig, in collaboration with Eric Atwell, Roger Garside and Geoffrey Leech. 1986. *The Tagged LOB Corpus: Users' Manual*. Bergen: Norwegian Computing Centre for the Humanities.

Johansson, S., G. N. Leech and H. Goodluck. 1978. *Manual of Information to Accompany the Lancaster-Oslo/Bergen Corpus of British English, for Use with Digital Computers*. Department of English, University of Oslo.

Knowles, Gerry and Lita Lawrence. 1987. 'Automatic Intonation

Assignment'. In Roger Garside, Geoffrey Leech and Geoffrey Sampson (eds.), *The Computational Analysis of English: A Corpus-based Approach*. London: Longman Group Ltd., pp. 139–148.

Meijs, Willem. 1982. 'Exploring Brown with QUERY'. In Stig Johansson (ed.), *Computer Corpora in English Language Research*. Bergen: Norwegian Computing Centre for the Humanities, pp. 49D65.

Sager, Naomi. 1981. *Natural Language Information Processing: A Computer Grammar of English and its Applications*. Don Mills, Ontario: Addison-Wesley Publishing Co.

Sampson, Geoffrey. 1987. 'Probabilistic Models of Analysis'. In Roger Garside, Geoffrey Leech and Geoffrey Sampson (eds.), *The Computational Analysis of English: A Corpus-based Approach*. London: Longman Group Ltd., pp. 16–29.

Sinclair, John (ed.). 1987. *Looking Up: An Account of the COBUILD Project in Lexical Computing*. London: Collins ELT.

Steen, Gert J. van der. 1984. 'On the Unification of Matching, Parsing and Retrieving in Text Corpora.' *ICAME News*, 8 (II): 41–46.

Steen, Gert J. van der. 1982. 'A Treatment of Queries in Large Text Corpora'. In Stig Johansson (ed.), *Computer Corpora in English Language Research*. Bergen: Norwegian Computing Centre for the Humanities, pp. 49–65.

Svartvik, Jan. 1987. 'Taking a New Look at Word Class Tags'. In Willem Meijs (ed.), *Corpus Linguistics and Beyond*. Amsterdam: Rodopi, pp. 33–44.

Taylor, Lita and Geoffrey Leech. 1989. Lancaster Preliminary Survey of Machine-Readable Language Corpora. Bergen, Norway: The Norwegian Computing Centre for the Humanities.

Reviews

Christian Mair. *Infinitival Complement Clauses in English. A Study of Syntax in Discourse.* Studies in English Language. Cambridge: Cambridge University Press, 1990. Pp. vii, 264. Reviewed by **Christer Geisler**, University of Uppsala.

This book marks the start of a new series, *Studies in English Language*, a series of monographs on English usage and English descriptive linguistics. Mair's study investigates infinitival clauses as complements to verbs, as in *They believed us to have sold the house*, and infinitives that function as sentence subjects, as in *To get pregnant was a rather stupid thing to do.* The book consists of four chapters, three appendices, and includes a detailed index.

Chapter 1, Introduction, states the aim of the study and presents the corpus-based material. The study is based on the entire material of the Survey of English Usage at University College London (179 spoken and written texts, and a total of 895,000 words running text). Terms such as 'complement' and 'complementation' are discussed, and the role of corpus-based studies is compared to formal syntactic frameworks. Chapter 2 discusses the syntax and the semantics of infinitival subject clauses and their extraposed variants, as well as subject clauses introduced by *for*, as in (1)–(3).

(1) *To disregard the past* is dangerous.
(2) It is dangerous *to disregard the past.*
(3) *For Beaminster to lose its sixth form* would be a tragedy for the town.

Chapter 3 investigates *to*-infinitival clauses as complements of transitive verbs. In this chapter postverbal complements of matrix predicates are analyzed. Mair focuses on four groups of syntactic complement patterns: monotransitive, ditransitive, complex transitive, and elliptical adverbial infinitives, as in (4)–(7).

(4) A grant enabled *them to complete the job.*
(5) She asked me to ask *the bank manager to read it through.*
(6) I consider *him to have excellent personal qualities.*
(7) I think his grandson uses *it to dig up worms with.*

Chapter 4 finally forms a summary and conclusion. In addition, there are three appendices, and an index which lists both individual matrix verbs and some of the terminology used.

Chapter 1 has no critical evaluation or definition of the key notions 'complementation' or 'complement', nor is the syntactic constituency of infinitives defined. On p. 2 Mair writes: "...I do not commit myself to any preconceived definitions of what is or what is not a 'complement'." As Huddleston 1988:350–52 points out, the terms complementation and complement are inadequately defined in Quirk et al 1985. The infinitive is regarded as clausal, but no mention is made of its structure, whether it has a subject at an underlying level of representation (for a discussion of the constituency of infinitives, see Koster and May 1982). Mair does not argue for the inclusion of infinitival clauses as sentence subjects among complement clauses (cf. Quirk et al 1985:1169n, which specifically states that subjects are not included in complementation).

Mair attacks generative grammar and formal syntactic frameworks in several places. For instance, he claims that '...the systemic idealisation and formalisation of the empirical data is a method which has outlived its usefulness...' (p. 219). Whether this is done in order to substantiate an atheoretical standpoint or to voice a disagreement against work on formal syntax is not clear. What is surprising, however, is that Mair frequently uses concepts that originate in formal syntax, such as the notion of control (p. 106) and various movement rules (eg Tough-movement in chapter 2, and various types of raising especially in chapter 3). Mair's criticism of formal grammar seems out of place and does not support his argumentation on other matters.

In chapter 2 subject infinitives are discussed; Mair finds extraposed infinitives to be the unmarked form. When infinitival clauses occur in subject position, they generally contain given information and have anaphoric relations to previous discourse (cf. the notion of 'proposition-linking' in Fox and Thompson 1990:301). It has been claimed that infinitival clauses in subject position express 'potentiality' (but although Mair refers to various works on this subject Perkins 1983 is not cited). Mair argues that the question of modality in infinitival subject clauses depends on the position of the infinitive and presence or absence of a *for*-subject (pp. 48–52, 84–92). Non-extraposed infinitives without *for* do not convey modality when compared with *-ing*-clauses in the same position. However, *for*-subject clauses typically express theoretical modality (cf. Allerton 1988). Mair discusses the status of *for* and claims that it is sometimes to be regarded as a preposition and not a subordinator (pp. 22, 40ff, and cf. Chomsky 1986, who terms it a prepositional complementizer).

Chapter 3 analyzes the many problems involved in the pattern Verb+NP+infinitive. First, monotransitive complement clauses are dealt with, as in *They want us to leave*. Mair states that the second NP *us* is a 'raised' subject, ie raised from subject position in the non-finite clause to object in the matrix clause. This goes against the consensus view that there is in fact a clause boundary between *want* and *us*. Not even Quirk et al 1985 says anything about raising in this kind of pattern (cf. Söderlind 1971). It is not clear to me how matrix verbs that allow passivization can be termed monotransitive, as in eg *Mary was expected to arrive any minute*. On pp. 97, 100–101 Mair discusses some of the inconsistencies of classification of verb complementation of Quirk et al 1985 (cf. Allerton 1988 and Huddleston 1988). Especially Quirk et al's complex transitive group is reorganized so that verbs which they include in this group are classified as monotransitive verbs. On p. 106 the notion of 'control', a theoretical concept concerning the way in which an NP is coreferential with the implied subject of an infinitive, is introduced, but should perhaps have been mentioned earlier in the study.

Mair's major contribution in chapter 3 is his reclassification of several verbs in Quirk et al's SVOC-class (the complex transitive type) and his discussion of adverbial infinitives in the SVOA and the SVO(A) groups. Among the verbs that have been misclassified according to Mair is *cause* (p. 101). The inherent modality of infinitival complement clauses is discussed in several places (pp. 102–103). Mair wisely avoids 'modality' in chapter 3 and opts for the term 'forward-looking'. Not only is Mair careful to distinguish infinitival complement clauses from adverbial infinitives, but he also considers the structural overlap with infinitival relatives (see eg pp. 72–75, 216–17).

Mair discusses the passive *be said to VP* constructions and claims that they are used to redistribute sentence information (p 180). This is, in my opinion, a predominant functional property of movement rules in terms of their importance for information flow. I wish Mair could have devoted more space to the given/new and the end-weight/end-focus dichotomies. Since subjects generally contain given information, it would have been interesting to see how NP-positions correlate with theories of information flow (for a recent study see Fox and Thompson 1990).

Mair also investigates some of the differences between subject-oriented (*in order to*-clauses: *Mary left to catch the train*) and object-oriented purpose clauses (*Mary bought a book to read*). The term 'fused constructions' is defined on p. 110 and reappears in the final chapter on p. 222 and forms the basis for a lengthy argumentation. But no instances are given in chapter 4 of the various movement rules that make up fused constructions, which makes the discussion difficult to follow.

The term 'statistically' is used throughout the study although no real statistical analyses are applied. Instead, differences in proportions between samples are expressed as numbers of occurrences, for example 52:5 in written versus spoken language, where straight percentages would have been more helpful (see eg pp. 112, 153, 156, 159). In addition, sometimes several parameters are discussed, as in the case with ditransitive complements in speech and writing (p. 153). First we are given the score 296:212 which represents the totals of ditransitive patterns in written and spoken texts, respectively. Then the ratio of active matrix verbs in written and spoken texts is given as 205:158, followed by the ratio of passive matrix verbs 91:54. This is difficult to interpret, and a simple crosstabulation including percentages would be more informative for the reader:

Ditransitive patterns (Mair, p. 153)

	Writing	Speech	TOTAL	%
ACTIVE	**205**	**158**	**363**	71%
	56%	44%		
PASSIVE	**91**	**54**	**145**	29%
	63%	37%		
TOTAL	**296**	**212**	**508**	
%	58%	42%		100%

The index lists mainly matrix verbs and to a lesser extent terminology. Entries which should have been included are 'modality in infinitives', 'relative infinitives', and perhaps also a list of verbs that are reanalyzed compared with Quirk et al 1985.

Chapter 2 on subject infinitives as well as chapter 3 on especially adverbial infinitives of the SVOA and SVO(A) patterns and their affinity with verb complements and infinitival relatives are the strongest parts of the book, because they pay attention to previously neglected topics in English linguistics. Chapter 3 which deals with transitive complementation is important because of Mair's reclassification of several verbs in Quirk et al 1985 (chapter 16). However, what all three chapters lack is detailed summaries and conclusions at the end of each chapter. I would also have liked to see short summaries in each of the four sections in chapter 3, ie sections 3.2–3.5. The concluding chapter is very short and does not fully summarize all sections of the book. It is unfortunate that chapter 4 contains no examples to support the argumentation and to illustrate what Mair terms 'fused' constructions.

Mair has presented us with a work full of interesting details about infinitival constructions in English. The book is a good example of qualitative research on corpus-based data at the Survey of English Usage. It is particularly stimulating to read about the syntax of infinitives in spoken discourse. The study benefits from an argumentative tone throughout, because Mair argues well for his analyses and defends particular claims in a convincing way. Furthermore, Mair calls attention to a number of marginal constructions such as different kinds of adverbial infinitival clauses and their affinity with other infinitive complement constructions.

References

Allerton, David J. 1988. Infinitivitis in English. In *Essays on the English language and applied linguistics on the occasion of Gerhard Nickel's 60th birthday*, edited by Josef Klegraf and Dietrich Nehls, 11–23. Heidelberg: Julius Groos.

Chomsky, Noam. 1986. *Lectures on government and binding*. 4th ed. Dordrecht: Foris.

Fox, Barbara A. and Sandra A. Thompson. 1990. A discourse explanation of the grammar of relative clauses in English conversation. *Language* 66. 297–316.

Huddleston, Rodney. 1988. Review of *A comprehensive grammar of the English language*, by Randolph Quirk, Sidney Greenbaum, Geoffrey Leech, and Jan Svartvik. *Language* 64. 345–354.

Koster, Jan and Robert May. 1982. On the constituency of infinitives. *Language* 58. 116–143.

Perkins, Michael. 1983. *Modal expressions in English*. London: Frances Pinter.

Quirk, Randolph, Sidney Greenbaum, Geoffrey Leech, and Jan Svartvik. 1985. *A comprehensive grammar of the English language*. London: Longman.

Söderlind, Johannes. 1971. Infinitive analysis: A comparison between different analytical approaches. *English Studies Today*. 5th Series. Istanbul. 123–134.

Jan Svartvik (ed.). *The London-Lund Corpus of Spoken English*. Lund University Press 1990. 359pp. Reviewed by **Jane A. Edwards**, University of California, Berkeley.

This volume is divided into two parts. Part I (59 pp) gives details of the London-Lund Corpus and its history, and lists over 200 published works known to have used LLC data in examining properties of spoken language. Part II (270 pp) reports results of the 'Text Segmentation for Speech' (TESS) project in Lund. The main goals of this project were to increase understanding of prosodic, syntactic and lexical properties of spoken language, and to apply this knowledge in algorithms for use in improved text-to-speech conversion.

As Part I indicates, the complete LLC consists of 100 texts each of 5000 words of spoken educated British English, covering a range of discourse types (monologue and multi-party interaction; spontaneous and prepared speech; formal and informal; covertly- and overtly-recorded), and produced by speakers of various ages and occupations and in differing relationships to one another. The utterances in the LLC are segmented into prosodically defined chunks, called 'tone units'. A tone unit is defined 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" (p73). The data are also systematically marked for points of pitch prominence ('nuclei'), direction of nuclear tones, lengths of pauses (two categories), and levels of stress (two categories). Thanks to its wide sampling of discourse types and its rich prosodic coding, the LLC is a valuable resource for a wide range of research questions. The 200-item bibliography cites both European and American research using the LLC, and will be of interest both to those using the LLC and to those interested in corpus-based results concerning particular topics.

Part II begins with a chapter by Jan Svartvik, providing background information concerning the TESS project. He compares the LLC with two written corpora (LOB and Brown) with respect to the relative lengths of their basic units and their most frequent lexical items. The sentences of the written texts are reported to average 18 or 19 words, whereas the tone units of the LLC average only 4.5 words; a number of words which are common in spoken language are rare in written language. Svartvik suggests that the poor quality of past speech synthesis resulted in part from using the syntactically- or graphemically-defined *sentence* as the primary unit of segmentation rather than the shorter units (i.e., information or tone units) more typical of speech. In addition, he suggests that quality might be improved by greater understanding of relationships between prosodics and grammar.

The main goals of TESS were to identify prosodic, syntactic and lexical regularities in spoken language and to use them in constructing algorithms for automatically segmenting written text into tone units for improved text-to-speech conversion. In order to describe these regularities, the project utilized an especially fine-grained set of 200 word class tags. This tagset is described in Chapter 3 by Svartvik. It includes semantic subclasses for several major syntactic categories (e.g., separate tags for adverbials of process, state, time, degree, manner, and reason), and separate tags for items serving discourse functions (e.g., 'well', 'you know'), which do not fit well into standard syntactic categories. Another property of tagging in TESS is that collocations, such as 'by the way', are tagged as unified wholes, rather than being analyzed into parts. These are clearly desireable properties in describing spoken language and are not found in all tagging systems. In addition to word class, five types of phrase were distinguished: noun phrase, verb phrase, prepositional phrase, adverbial phrase, and adjectival phrase. Syntactic roles such as subject and object were not tagged explicitly in TESS, but rather were to be inferred on the basis of positional and other information.

Chapter 4, by Mats Eeg-Olofsson, describes an automatic word class tagger. When applied to three (written language) editorials from the Brown Corpus, it had an accuracy of 94–97%. The chapter also describes a prototype phrase parser, which bases its categorizations on co-occurrence probabilities of word class tags in other texts, together with a general strategy of delaying boundaries as long as possible (p119). Both the tagger and phrase parser were designed for coding efficiency and are not intended as models of online human language processing. Chapter 2 contains some discussion of how these processes may diverge.

The remaining chapters can be divided roughly into two content areas: linguistic and computational linguistic.

In the linguistic category, the principal topics addressed are: (1) semantic, syntactic and prosodic properties of particular discourse items and adverbials (Chapters 5 and 9 by Anna-Brita Stenström and Chapter 6 by Bengt Altenberg), (2) semantic functions of the prosodic feature known as the 'booster', that is, a sudden shift in pitch (Chapter 7 by Altenberg), (3) the statistical distribution of filled and unfilled pauses and various discourse items with respect to tone units, clauses, phrases and other units within dialogue and monologue (Chapter 8 by Stenström), (4) the limited value of adverbial-plus-comma sequences in written texts for predicting the presence of tone unit boundaries next to the same adverbials in spoken language (Chapter 9 by Stenström). These are summarized briefly below.

Concerning the first topic, a recurrent theme in the book is that the same lexical item may serve different functions depending on its prosodics and its position in the utterance (initial, medial, final). In addition, it may have a different primary function in spoken compared to written language (e.g., 'now' as a narrative continuation marker, or a temporal adverb). Chapter 5 contains a fine-grained prosodic, positional, semantic, and syntactic analysis of items which serve discourse functions, such as 'well', 'you know', and 'please'. The chapter provides an especially close look at 'really', and the so-called 'right' set ('right', 'alright', 'okay', 'yeah', 'mhm', etc.). Chapters 6 and 9 provide similar analyses, but for adverbials. In addition, Chapter 6 compares two learner's dictionaries, COBUILD and LDOCE, with reference to the degree to which they incorporate information concerning spoken language into their entries. Altenberg notes that both dictionaries are based on corpus data, but that only COBUILD uses frequency of use as a principle for ordering entries. He gives good arguments supporting the need in such dictionaries for a more detailed typology of adverbial functions, and for more systematic inclusion of prosodic and positional information for discriminating among the different senses and uses of a word.

Concerning the second topic, Altenberg surveys the functions served by pre-nuclear boosters in five texts from the LLC (including a monologue lecture). These include: intensifying quality and quantity ("we met *all* the American Chairmen"), emphasizing truth and modality ("he *certainly* stirred the place up"), and emphasizing contrasts ("*only* one"). In addition, the booster and nucleus sometimes jointly serve a bracketing function, marking the beginning and end of "a group of words that belong together semantically and syntactically" (p. 205).

Concerning the third topic, in Chapter 8, Stenström examines the distribution of silent pauses, filled pauses (e.g., 'e:m') and 'verbal fillers' (e.g., 'well', 'you know') in spontaneous dialogues and in a monologue lecture (S.12.6). Some of the distributional findings are based on such small frequencies that one wonders whether they would be statistically significant. The following findings, however, are quite robust. In both dialogue and monologue texts, filled pauses occur more often *within* (70%) than *between* tone units (30%), while the reverse is true for silent pauses (80% between; 20% within). Also, contrary to some expectations from the literature, silent pauses were found to occur more often *within* clauses (e.g., between subject and verb, between verb and object, etc.) than *between* them (39% vs. 11%). This chapter is difficult to read in places. Some headings and labels in Tables and Figures are not defined (e.g., 'combination' in Table 8:1 may or may not include verbal fillers), and some subsection summaries seem to contradict the data they summarize. For example, on p. 243 Table 8:12 is said to show that "brief FPs were always found within the tone unit," while the table itself shows some brief FPs also *between* tone units.

Concerning the fourth topic, in Chapter 9 Stenström assesses the degree to which an adverbial in speech is likely to co-occur with a tone unit boundary in speech, given that it co-occurs with a comma in writing. Predictability is found to be best for adverbials in utterance-initial positions. For those later in the utterance, predictability is somewhat improved by rules using probabilities of co-occurrence of particular functions and prosodic contours as extracted from other LLC data, but remains somewhat limited.

Concerning the computational linguistic findings of the book, Chapters 11, 12, and 13 present the main results on the automatic segmentation of written text into tone units. Chapter 11 by Altenberg summarizes a set of prosodic rules useful in describing and predicting the location of tone unit boundaries within a monologue

(S.12.6). This text is a lecture, partly pre-planned, and possibly partly read, and was chosen because of its nearness to written language. Altenberg observes that predictability of tone unit boundaries is greatest between clauses (compared to between elements within clauses and between phrases), but that the degree of predictability depends on clause type. For example, tone unit boundaries are found 100% of the time after initial clauses, but only 33% of the time before non-finite nominal clauses. Predictability is improved by a rule which involved adding a tone unit boundary immediately after a matrix clause with SVC, SVO or SVA structure except when the matrix clause is shorter than five words (p281). Between-clause tone unit boundaries could be predicted 95% of time. In contrast, within-clause tone unit boundaries are more variable and may resist prediction by means of lexicogrammatical rules, since they depend on higher level discourse structures and semantic processes such as those related to location of informational focus.

Chapter 12, by Altenberg, presents an algorithm constructed from the descriptive rules of Chapter 11, and reports on a test of it involving segmentation of a written editorial from the Brown Corpus. The algorithm involves applying 11 rules, divided into two cycles. The first cycle consists of five rules intended to assign tone unit boundaries at major syntactic boundaries; the second cycle involves seven rules for minor syntactic boundaries. The algorithm proceeds in an inchworm-like manner. Once a tone unit boundary has been assigned by a rule in the first cycle, the algorithm back-tracks to the beginning of that unit (often a clause) and looks within it for structures satisfying one or more of the rules pertaining to minor syntactic boundaries. One major syntactic rule involves replacing all major punctuation marks with tone units boundaries, except for comma before such phrases as 'for instance'. Others refer to particular word classes (e.g., coordinating conjunctions), and are sometimes modified by position (initial, medial, etc.) or distance from the preceding or following tone unit boundary. Finally, there are some general constraints, such as one preventing a tone unit boundary following a 'light' first element (e.g., pronoun or closed-class adverb). This constraint seems similar to one proposed in Gee and Grosjean (1983), who were, however, concerned with segmentation into pause-bounded units or rhythmically-defined units rather than tone units. A fertile area for future work would be to compare the rules of these two systems in greater detail as a basis for further insight into the structure of these various types of units in speech. In any case, to conclude this summary of the TESS segmentation, when the TESS segmentation algorithm was applied to the tagged editorial from Brown, it succeeded in predicting 90% of the tone unit boundaries which would be found when it was read aloud. Chapter 13, by Eeg-Olofsson, describes the Prolog implementation of the algorithm.

Chapter 10, by Svartvik describes a menu-driven system for automatically converting the prosodic distinctions of the LLC data into two types of visual displays: Bolinger's squiggly diagrams and variably elevated plateaus representing relative prominence of syllables. *This offers improved visualization of trends in the data.*

This book is an important contributition to corpus based research. Like the Garside, Leech & Sampson (1987) volume, reviewed in last year's *ICAME Journal*, the Svartvik volume presents results of an extended research program focussing on a large corpus of connected texts of actual language in use. In contrast to Garside et al., however, its primary focus is on spoken language, and especially the interaction of prosodics and grammar.

Especially valuable in this book are its rich descriptions of prosodic, lexical and syntactic properties of adverbials and discourse items, its list of lexico-syntactic rules useful in predicting prosodic segmentation of spoken texts, and its report on the successful automatic segmentation of written language into tone units. This type of work is complementary to experimental work involving prosodics in language understanding. As one example, Price et al. (1991) presented listeners with sentences which are similar phonetically but different prosodically and reflect a wide range of different structural types (e.g., "Gary knew we were worried, but he'd lied" compared to "Gary knew we were worried Buddy'd lied"). They then performed careful acoustic analyses to help determine which aspects of prosodics listeners relied upon the most in distinguishing between the paired alternatives. In another vein, researchers concerned with models of lexical access and moment-to-moment language understanding are also increasingly emphasizing the importance of prosodics (e.g., Cutler, 1989; Frauenfelder & Lahiri, 1989). Hopefully, much more work will continue using a variety of methods but with a common focus on relationships between prosodics and grammar in spoken language processing.

I would recommend this book very highly to all researchers interested in corpus-based linguistics and speech synthesis. In addition, I would encourage people not yet interested in corpus linguistics to take a look at this book as a partial introduction to the area. It is impressive in the diversity and high quality of its contributions, and in the creative use of corpus data in expanding our knowledge of prosodic, lexical and syntactic aspects of spoken language.

References

Bachenko, J. & E. Fitzpatrick. 1990. A computational grammar of discourse-neutral prosodic phrasing in English. *Computational Linguistics*, 16, 155–170.

Cutler, A. 1989. Auditory lexical access: Where do we start? In W. Marslen-Wilson (ed.) *Lexical Representation and Process*. Cambridge, Mass.: MIT Press.

Frauenfelder, U. & A. Lahiri. 1989. Understanding words and word recognition: Can phonology help? In W. Marslen-Wilson (ed.) *Lexical Representation and Process*. Cambridge, Mass.: MIT Press.

Garside, R., G. Leech, & G. Sampson (eds.). 1987. *The Computational Analysis of English: A Corpus-based Approach*. London: Longman.

Gee, J. P., & F. Grosjean. 1983. Performance structures: A psycholinguistic and linguistic appraisal. *Cognitive Psychology*, 15, 411–458.

Price, P. J., M. Ostendorf, S. Shattuck-Hufnagel, & C. Fong. 1991. The use of prosody in syntactic disambiguation. Stanford Research Institute, manuscript.

Magnus Ljung. *A Study of TEFL Vocabulary.* Acta Universitatis Stockholmiensis. Stockholm Studies in English LXXVIII, 1990. 425pp. Reviewed by **Felicitas Tesch,** Technische Universität Berlin.

This book is divided into two main parts, the first containing a critical analysis of the vocabulary of fifty-six Swedish textbooks for the upper secondary schools. The second part consists of detailed word lists, which illustrate the evaluation discussed in the first part.

Generally, this book supplies information of three kinds: First, the structure of the corpus is presented in detail. The vocabulary of the textbooks is then compared to the findings of the corpus of authentic English. Finally, the word lists of the textbook vocabulary provide useful information both for linguists, teachers, and textbook compilers.

In the introduction the different textbooks chosen for the study – viz. the GYM-corpus – are explained. There are some minor problems in the GYM-corpus: The textbooks differ considerably in length; some are not approved by the Swedish education authorities and some cannot be assigned with much exactitude to a certain type of school. But on the whole, the GYM-corpus, comprising almost one and a half million words, is a valuable basis of comparison. The type-token ratio can be applied to corpora of 1.5 million running words of written text.

Although the author's arguments for not having chosen the LOB or the BROWN corpus for comparison are not too evident, the selection of the COBUILD corpus can be accepted, because this is the largest one. Definitely more words are needed for lexical than grammatical analysis.

In the evaluation, the two corpora presented above are compared. The comparison starts with listing the unique words in GYM and COBUILD. The main difference lies in the abstract-concrete distinction. Whereas the unique words occurring in the GYM-corpus mostly designate physical objects or actions, the majority of the unique COBUILD words are abstract nouns and verbs denoting interpretations of actions. As 20% of the thousand most frequent words in one corpus are not found in the other, it can be stated that there is quite a great variation due to type of text. The same results apply to the shared words: There is a clear overrepresentation of words denoting concrete phenomena in the GYM-corpus. In addition, it could be shown that the textbooks prefer certain topics like *school*, *home*, and *England/America*, topics that make these words more likely to occur in the texts. There are also stylistic differences between the GYM and the COBUILD corpus. Generally, the TEFL-texts prefer more informal words like *mum* and *dad* (p.18), as well as contracted forms.

Apart from the comparison of individual words, Ljung has also attempted to characterize in general the communicative profile of the texts, i.e. text-type comparison. On this level, he includes the London-Lund Corpus of spoken English. On the dimension ranging from abstract content to situated reference, the GYM texts are in an intermediate position between the written and the spoken authentic corpora. From the number of past-tense forms, the author concludes that there are more narrative texts in the GYM-corpus than in the other two corpora.

The comparison between the GYM-corpus and the London-Lund corpus could be carried out only at word level. The author here presents one interesting finding: The GYM texts contain more what he calls "phrasal verbs" (e.g. *take off*) than both COBUILD and the London-Lund corpora.

The author also sheds some light on cultural differences in the corpora. There is a clear orientation in favour of British English in both corpora. As far as the use of words denoting nationality is concerned, the British dominate in both corpora, followed by the Americans. Not surprisingly, the textbooks add the Swedes.

At the end of the evaluation, some characteristics of the subcorpora are mentioned.

To sum up, two points need some further discussion:

As the author himself points out, the "comparison corpus [i.e. COBUILD, FT.] ... has to be machine-readable" (p.4); the same criterion should be applied to the word lists in the GYM-corpus. If they were presented in machine-readable form, they would be of even more value to other investigators.

Second, there remain some problems of lemmatization. The author mentions the ambiguity of a word like *can* (p.5), which can easily be made clear by its context. On the other hand, word-class problems (e.g. prepositions or adverbs) are not discussed here and need some further investigation.

Generally, this study provides interesting findings in the comparison of a textbook corpus with a corpus of authentic English. It gives insight into the text-type characteristics of Swedish textbooks for English and is of great value both for linguists, textbook authors, and teachers of English as a foreign language.

Eva Leitzke, *(De)nominale Adjektive im heutigen Englisch: Untersuchungen zur Morphologie, Syntax, Semantik und Pragmatik von Adjektiv-Nomen-Kombinationen der Typen* atomic energy *und* criminal lawyer. Linguistische Arbeiten 221. Tübingen: Niemeyer, 1989. 200 pp. ISBN 3–484–30221–6. Reviewed by **Christian Mair**, University of Freiburg.

This book – a slightly revised version of the author's 1988 Munich Ph.D. thesis and, dare I say it, unfortunately written in German – is a study of adjectives derived from nouns with the suffixes *-al*, *-an*, *-ary*, *-ic*, *-ine*, *-ish*, *-ly*, *-ous*, *-y* mainly based on data from the LOB microfiche concordance and corpus text. The author deplores the fact that the vicissitudes of having to run the computer-readable tagged LOB on a mainframe made the full statistical evaluation of the corpus evidence impossible for her (p.6). That ICAME's current offer – LOB for use on PCs and TACT as an easy-to-use retrieval program (generously made available by the developers at the University of Toronto) – would easily have taken care of most of her problems is a good indication of the progress made in corpus development over the past few years and of the potential of ICAME's basic concept, i.e. supplying individual scholars who are not necessarily computer experts with the tools for decentralized linguistic research.

That the function of, say, the adjective *musical* is different in *a musical instrument* from *a musical child* has, of course, long been recognized and frequently been commented on in the linguistic literature. Here, as in many similar studies, the corpus serves as a useful corrective in the "academic" (notice the vagueness of the use) debate, because it allows the analyst to exclude infrequent or untypical usage from consideration. This leads to a realistic and balanced appraisal both of the desirable but often unsupported generalizations and the much discussed but empirically unattested "crucial counter-examples" the discussion of which often takes up reams of paper without a solution coming any nearer. It is, incidentally, one of the strengths of the present work that Leitzke not only reviews literature in English linguistics but also work done on related phenomena in the Romance languages and German.

After a meticulous discussion of the morphological, syntactic, semantic and pragmatic properties of these adjectives, Leitzke posits a gradient extending from prototypical nouns through purely "relational" adjectives, such as *atomic* in *atomic bomb*, which are adjectives morphologically but behave like nouns in almost all other respects, to the prototypically adjectival "qualifying" adjectives (e.g. *industrious* in *industrious pupil*).

The main merit of Leitzke's study – and an impressive justification of the gradient model assumed – is the thorough treatment of the borderline area between relational and qualifying adjectives. Thereby, the evidence from the LOB Corpus serves to dispel wishful thinking on the part of linguists concerning, for example, the systematicity of suffix variation. Except for the pair *historic/historical*, where the former is indeed relational and the latter qualifying, the match between form and meaning is never perfect. In chapter IV, dealing with polysemous adjectives like *organic* or *peripheral*, which can be relational or qualifying, examples from LOB are used to illustrate the syntactic and semantic factors determining the meaning of an item in text. Etymologically "foreign" suffixes, attributive use, modification by restrictive rather than intensifying adverbs (compare "strictly military" to "very military") and a number of semantic and contextual clues which for reasons of space cannot be illustrated here are shown to favour the interpretation of a given polysemous adjective as relational rather than qualifying. A list of all the relevant types attested in LOB concludes the book.

Shorter notices

Corpus Studies in Japan

Makoto Shimizu
Kyushu Institute of Technology

Since the advent of personal computers in the 1980s, more and more Japanese linguists have started using computers and computer corpora in their research. The following is a bibliography of papers written by Japanese English linguists whose topics are, more or less, relevant to computer corpora, especially the LOB Corpus and the Brown Corpus. Most of the authors are either *ICAME Journal* readers or have obtained computer corpora through ICAME, or both.

Akano, Ichiro (1990a) "Invitation to corpus linguistics (1): Brown Corpus" *Sell* 6, 142–8.

Akano, Ichiro (1990b) "Invitation to corpus linguistics (2): What is OCP?" *Academic Bulletin of Kyoto University of Foreign Studies* 35, 1–15.

Fukamachi, Jun (1990) "A retrieval system of tagged corpora with multi-conditions" *Bulletin of Language Centre*, Nagoya University 2, 199–211.

Fukushima, Naoyuki (1991) "An estimate of the syllable types in English with the use of computerized dictionary" *Walpurgis* (Bulletin of Kokugakuin University).

Hojo, Kazuaki (1988) "The semantics and usage of preterite subjunctive patterns in present-day English with special emphasis on *as if* clauses" *Acta Humanistica et Scientifica Universitatis*

Sangio Kyotiensis (Bulletin of Kyoto Sangyo University) Volume 18 Humanities Series No. 15, 150–91.

Kato, Kazuo (1986) "The difficulty of translation of dictionary definitions" *Annual Report of Iwate Medical University, School of Liberal Arts and Sciences* 21, 113–26.

Kato, Kazuo (1989) "Causative *have* and accent" *Annual Report of Iwate Medical University, School of Liberal Arts and Sciences* 24, 1–12.

Maruta, Tadao (1988) "On the adjectives derived from verbs in English and Japanese" Report to Yamagata University.

Nakamura, Junsaku (1984) "Accommodating a drama text in a personal data-base" *The Bulletin of the English Language Education Society of Shikoku* 5, 97–109.

Nakamura, Junsaku (1985a) "On the methodology of quantitative groupings of English texts" *JACET Bulletin* 16, 133–48.

Nakamura, Junsaku (1985b) "Some quantitative analysis of On Golden Pond revealed through a data-base management system" *Journal of Cultural and Social Science, College of General Education, University of Tokushima* 20, 133–48.

Nakamura, Junsaku (1986) "Classification of English texts by means of Hayashi's quantification method type III" *Journal of Cultural and Social Science, College of General Education, University of Tokushima* 21, 72–86.

Nakamura, Junsaku (1987) "Notes on the use of Hayashi's quantification method type III for classifying texts" *Journal of Cultural and Social Science, College of General Education, University of Tokushima* 22, 127–45.

Nakamura, Junsaku (1989a) "Creation of vocabulary frequency tables from the Brown Corpus" *Journal of Cultural and Social Science, College of General Education, University of Tokushima* 24, 171–82.

Nakamura, Junsaku (1989b) "A quantitative study on the use of personal pronouns in the Brown Corpus" *The Bulletin of the College English Education Society* 20, 51–70.

Nakamura, Junsaku (1990) "A study on the structure of the Brown Corpus based upon the distribution of grammatical tags" *Journal of Foreign Language and Literature, College of General Education, University of Tokushima* 1, 13–35.

Nakamura, Masanori (1988) "English texts processing by computers and English language education" *Bulletin of Kyoto Tachibana Women's University* 15, 79–92.

Ohmi, Kazuo (1985) "An automatic concordance compiler: Universal Concordance Program, Osaka" *Studies in Language and Culture* 11, Faculty of Language and Culture, Osaka University 189–218.

Shimizu, Makoto (1985) "The endophoric use of English demonstratives" *Bulletin of Graduate School, Seinan Gakuin University* 4, 73–96.

Shimizu, Makoto (1987) "Coreference of pronominals and anaphors in Functional Syntax" *Bulletin of Hiroshima Jogakuin College* 37, 75–95.

Shimizu, Makoto (1988) "The pragmatic aspect of anaphora by reflexives" *Bulletin of Hiroshima Jogakuin College* 38, 33–50.

Shimizu, Makoto (1990) "A DRS approach to reflexives" *The Bulletin of the Kyushu Institute of Technology* (Humanities and Social Science) 38, 35–57.

Tachi, Kiyotaka (1988) "English text retrieval system" *Network* 1, Information Centre, Fukui University 9–25.

Takahashi, Kiyoshi (1989) "On performance of words", *AVEC Annual Report* 4, Tokyo University of Foreign Studies 15–28.

Takahashi, Kiyoshi (1990) "Non-volitional *will* in conditional *if*-clauses" *Bulletin of Tokyo University of Foreign Studies* 40, 55–64.

Umeda, Iwao (1987) "Psychological predicates in English" *International Review of Applied Linguistics* 25, 91–101.

Generally speaking, these papers can be divided into four types; 1) Introductions to computer corpora, 2) manuals of software 3) quantitative studies, and 4) studies of usage or grammar.

1) These are meant to introduce Japanese linguists or language teachers to computer corpora. They include general guidance to computer corpora, often LOB and/or Brown, and the way to retrieve textual data from corpora. Some mention retrieval software. Papers of this type are Akano (1990a,b), Fukushima (1991), J. Nakamura (1984), M. Nakamura (1988), and Tachi (1988). Fukushima gives some information on Moby Pronunciator, an electric pronunciation dictionary compiled in the United States. J. Nakamura describes the process he input the English drama text *On the Golden Pond* written by Ernest Thompson.

2) Papers of this type are manuals of the software the authors have made. Fukuda (1990), Ohmi (1985), and Tachi (1988) are in this category. Fukuda's paper is about his text retrieving software designed for tagged corpora. Ohmi has developed a program called UCPO, based on the results of COCOA and OCP, which works on a mainframe computer. Tachi informs us about another text retrieving system running on Pascal.

3) Quantitative methods are used in papers of this kind. Fukushima (1991) discusses the method of predicting the number of syllables of an English word, using an electric pronunciation dictionary. Most of J. Nakamura's papers are concerned with detailed quantitative analyses of the words used in an English play.

4) The authors of this type of papers cite English texts fully rather than count the frequency of a certain word. Kato (1986) discuss the definition of lexical items. M. Nakamura (1988) and Takahashi (1989) examine the possibility of using computer corpora in the field of English language teaching. Hojo (1988) deals with the subjunctive patterns through data obtained from the Brown, LOB, and London-Lund corpora. Kato (1989) is about causative *have*. Maruta (1988) argues that English adjectives with the *-ing* form inherit their theta-roles from the base verb. Shimizu (1985) deals with English demonstratives. Shimizu (1987, 1988, 1990) try to refute GB Theory using data from the LOB, Brown and London-Lund corpora. Taka-

hashi (1990) is a study on non-volitional *will* in conditional *if*-clauses. Contrary to the claim by A.S. Hornby, Takahashi finds examples in which future-referring *will* is used in conditional clauses. Umeda (1987) discusses some aspects of psychological predicates such as *amuse*, *bore*, and *embarrass* with special reference to the co-occurrence of *very* and *by* plus (human) agents.

Most of the papers are written in Japanese. Papers written in English are all the papers of J. Nakamura, Hojo (1988), Umeda (1987), and Shimizu (1985). Other papers are in Japanese, but in each paper the examples retrieved from computer corpora are, of course, cited in English.

At the present stage, not too many studies have been made in the field of corpus linguistics in Japan. The number of personal computers, however, has risen drastically in this country, and there is a growing interest in computer corpora among Japanese linguists. Soon, I believe, this field will attract wide attention in this country as well.

Progress Report on the Text Encoding Initiative

Lou Burnard
Oxford University Computing Service

The first draft of the ALLC–ACH–ACL Text Encoding Initiative's recommendations for encoding of machine readable texts for interchange was published in July 1990, and reprinted with minor corrections in November. This volume, edited by C.M. Sperberg-McQueen and Lou Burnard, is the result of the first two years' work in the TEI project, which is due to produce its final report in June of 1992. A brief summary of the contents of the TEI's first draft was published in a recent issue of *Humanistiske Data* (3–90, pp. 52–58), and a longer account, which also discusses a few of its implications for corpus linguistics, is due to appear in the forthcoming Proceedings of the ICAME Conference held in Berlin last year.

The success of the Guidelines will be crucially dependent on the widest possible public discussion of their contents. Many individual comments have already been received, and it is hoped that readers of the initial draft will continue to send them in. Tutorial materials are also being prepared, to introduce the basic notions of the TEI to a wider audience. A first draft of an introductory guide "Living with the Guidelines" was recently presented at a well-attended TEI Workshop held in conjunction with the 1991 ACH–ALLC conference at Tempe, Arizona.

The major objective during the second TEI funding cycle will be to extend the scope and coverage of the Guidelines. Although the current draft report provides a good general framework for most forms of text-based scholarship, it is clear that many areas are only sketched out in it, and that some types of research have been barely covered at all. Two approaches are being taken to improve this situation. Firstly, the current recommendations will be put to the test by a wide variety of affiliated research projects. Secondly, a large number of specialised working groups are being set up to make specific recommendations for extensions and additions.

Affiliated projects

A number of major corpus-based research projects (for example, the British National Corpus, the Women Writers Project at Brown University, the Pandora Project at Harvard) are already affiliated with the TEI and will be working closely with it over the next few months. Following successful short workshops at Chicago and Tempe, the TEI is planning to hold further intensive workshops for affiliated projects to gain and exchange experience in using the Guidelines in a practical context, both in Europe and in North America.

Working groups

A large number of small but tightly-focussed working groups have already been set up to make recommendations in specified areas, either directly where an area is already well-defined, or indirectly by sketching out a problem domain and proposing other work groups which need to be set up within it. A list of the groups so far constituted follows:

TR1: Character sets (Harry Gaylord, University of Groningen)

TR2: Text criticism (Robert Kraft, University of Pennsylvania)

TR3: Hypertext and hypermedia (Steven De Rose, Electronic Book Technology)

TR4: Mathematical formulae and tables (Paul Ellison, University of Exeter)

TR6: Language corpora (Douglas Biber, Northern Arizona University)

TR8: Physical description of mss and incunabula (Jacqueline Hamesse, University of Louvain la Neuve)

TR9: Analytic bibliography of printed books (John Barnard, University of Leeds)

AI1: General linguistics (Terry Langendoen, University of Arizona)

AI2: Spoken texts (Stig Johansson, University of Oslo)

AI3: Literary studies (Paul Fortier, University of Manitoba)

AI4: Historical studies (Daniel Greenstein, University of Glasgow)

AI5: Machine-readable dictionaries (Robert Amsler, Mitre Corporation)

AI6: Computational lexica (Robert Ingria, BBN)

AI7: Terminological databases (Alan Melby, Brigham Young University)

Each group is formally assigned to one of the two major working committees of the TEI, depending on whether its work is primarily concerned with Text Representation (TR) or Text Analysis and Interpretation (AI), and it is the Working Committees which will be primarily responsible for reviewing and endorsing the results of the Work Groups as they become available over the rest of this year.

For more information, and up to date information about the progress of the TEI, please get in touch with either of the TEI editors, Michael Sperberg-McQueen (U35395@UICVM.BITNET) or Lou Burnard (LOU@UK.AC.OXFORD.VAX), or subscribe to the public discussion list TEI–L@UICVM.BITNET.

Eleventh ICAME Conference

The 11th ICAME Conference on English Language Research on Computerized Corpora was held in Berlin, 10–13 June, 1990. Some thirty-five papers were read on a variety of aspects of English corpus work: corpus design and editing, text processing and retrieval, description of English (grammar, collocations, varieties, pragmatics). A report on the conference will appear later this year in the *CCE Newsletter* (*Computer Corpora des Englischen in Forschung, Lehre und Anwendungen*, ed. by Gerhard Leitner). Papers from the conference will be published in a volume edited by Gerhard Leitner.

The participants are indebted to the organisers (Gerhard Leitner and his team) for a successful and well-organised conference. The next ICAME conference will be held at Leeds in May 1991.

The ICAME network server

A network server has been set up at the EARN/BITNET node in Bergen (coordinator: Knut Hofland). The server can be reached from any network that has a gateway to EARN/BITNET like Uninett, Janet, Internet, Csnet, etc. The server holds information about the material available, some text samples, order forms, an ICAME bibliography, a survey of text corpora, programs and documentation, and network addresses. See further *ICAME Journal* 12 (1988), pp. 81–83.

Send a note with subject: DIR to get more information.

Address server

EARN/BITNET:	FAFSRV@NOBERGEN
JANET:	FAFSRV@EARN.NOBERGEN
INTERNET:	FAFSRV%NOBERGEN.BITNET@CUNYVM.CUNY.EDU

Address coordinator

EARN/BITNET:	FAFKH@NOBERGEN
JANET:	FAFKH@EARN.NOBERGEN
INTERNET:	FAFKH%NOBERGEN.BITNET@CUNYVM.CUNY. EDU

Another file server at the Norwegian Computing Centre for the Humanities

The machine nora.navf-edb-h.uib.no has been established as a mail based server for the Norwegian Computing Centre for the Humanities. Information is grouped in different catalogues, some of which have information only in Norwegian. The relevant catalogues for ICAME are icame, ncch, info, pc, mac, and unix.

The server is called NAVFSERV and runs the DECWRL archive server. This server is more robust than FAFSRV. The material on FAFSRV has been transferred to NAVFSERV, and FAFSRV will gradually be phased out. NAVFSERV accepts three types of commands, and several commands can be placed in the body of the mail message. However, the results will be sent in one file, so do not request several large files in one message. The commands (can be sent in the Subject line or body):

Help	Help file
Index	Top level index
Index <catalogue>	Index for a catalogue
send <catalogue> <filename>	Fetch a file in a catalogue

Example: We want to get the files icame.cond and icame.material in the catalogue icame. Send the following note:

```
To: navfserv@nora.navf–edb–h.uib.no
Subject: whatever (or a command)

send icame icame.cond
send icame icame.material
```

Note that the file names are given in the form icame.cond in stead of ICAME COND as on the FAFSRV server.

The files are also available via anonymous FTP from nora.navf–edb–h.uib.no (129.177.24.42).

New material

There are plans to produce a CD-ROM with some of the corpora distributed through ICAME. See the leaflet accompaning the journal. The example text (transcription and audio cassette) for the London-Lund Corpus announced earlier is withheld due to production problems. As regards the new, edited version of the Polytechnic of Wales Corpus, see the presentation earlier in this issue.

Material available through ICAME

The following material is currently available through the International Computer Archive of Modern English (ICAME):

Brown Corpus, untagged text format I (available on tape or diskette): A revised version of the Brown Corpus with upper- and lower-case letters and other features which reduce the need for special codes and make the material more easily readable. A number of errors found during the tagging of the corpus have been corrected. Typographical information is preserved; the same line division is used as in the original version from Brown University except that words at the end of the line are never divided.

Brown Corpus, untagged text format II (tape or diskette): This version is identical to text format I, but typographical information is reduced and the line division is new.

Brown Corpus, KWIC concordance (tape or microfiche): A complete concordance for all the words in the corpus, including word statistics showing the distribution in text samples and genre categories. The microfiche set includes the complete text of the corpus.

Brown Corpus, WordCruncher version (diskette): This is an indexed version of the Brown Corpus. It can only be used with WordCruncher. See the article by Randall Jones, *ICAME Journal* 11, pp. 44–47.

LOB Corpus, untagged version, text (tape or diskette): The LOB Corpus is a British English counterpart of the Brown Corpus. It contains approximately a million words of printed text (500 text samples of about 2,000 words). The text of the LOB Corpus is not available on microfiche.

LOB Corpus, untagged version, KWIC concordance (tape or microfiche): A complete concordance for all the words in the corpus. It includes word statistics for both the LOB Corpus and the Brown Corpus, showing the distribution in text samples and genre categories for both corpora.

LOB Corpus, tagged version, horizontal format (tape or diskette): A running text where each word is followed immediately by a word-class tag (number of different tags: 134).

LOB Corpus, tagged version, vertical format (available on tape only): Each word is on a separate line, together with its tag, a reference number, and some additional information (indicating whether the word is part of a heading, a naming expression, a quotation, etc).

LOB Corpus, tagged version, KWIC concordance (tape or microfiche): A complete concordance for all the words in the corpus, sorted by key word and tag. At the beginning of each graphic word there is a frequency survey giving the following information: (1) total frequency of each tag found with the word, (2) relative frequency of each tag, and (3) absolute and relative frequencies of each tag in the individual text categories.

LOB Corpus, WordCruncher version (diskette): This is an indexed version of the tagged LOB Corpus (horizontal format). It can only be used with WordCruncher.

London-Lund Corpus, text, original version (computer tape or diskette): The London-Lund Corpus contains samples of educated spoken British English, in orthographic transcription with detailed prosodic marking. It consists of 87 'texts', each of some 5,000 running words. The text categories represented are spontaneous conversation, spontaneous commentary, spontaneous and prepared oration, etc.

London-Lund Corpus, KWIC concordance I (computer tape): A complete concordance for the 34 texts representing spontaneous, surreptitiously recorded conversation (text categories 1–3), made available both in computerized and printed form (J. Svartvik and R. Quirk (eds.) *A Corpus of English Conversation*, Lund Studies in English 56, Lund: C.W.K. Gleerup, 1980).

London-Lund Corpus, KWIC concordance II (computer tape): A complete concordance for the remaining 53 texts of the original London-Lund Corpus (text categories 4–12).

London-Lund Corpus, supplement (diskette): The remaining 13 texts of the 100 spoken texts collected and transcribed at the Survey of English Usage, University College London. See the presentation by Sidney Greenbaum, *ICAME Journal* 14 (1990) pp. 108–110.

Melbourne-Surrey Corpus (tape or diskette): 100,000 words of Australian newspaper texts (see the article by Ahmad and Corbett, *ICAME Journal* 11, pp. 39–43).

Kolhapur Corpus (tape or diskette): A million-word corpus of printed Indian English texts. See the article by S.V. Shastri, *ICAME Journal* 12, pp. 15–26.

Lancaster/IBM Spoken English Corpus (tape or diskette): A corpus of approximately 52,000 words of contemporary spoken British English. The material is available in orthographic and prosodic transcription and in two versions with grammatical tagging (like those for the LOB Corpus). There is an accompanying manual. See further *ICAME Journal* 12, pp. 76–77.

Polytechnic of Wales Corpus (tape or diskette): Orthographic transcriptions of some 61,000 words of child language data. The corpus is parsed according to Hallidayan systemic-functional grammar. There is no prosodic information. See further *ICAME Journal* 13 (1989), p. 20ff, and the presentation of the edited version of the corpus in this issue.

Most of the material has been described in greater detail in previous issues of our journal. Prices and technical specifications are given on the order forms which accompany the journal. *Note that tagged versions of the Brown Corpus cannot be obtained through ICAME. The same applies to audio tapes for the London-Lund Corpus, the Lancaster/IBM Spoken English Corpus, and the Polytechnic of Wales Corpus.*

There are available printed manuals for the LOB Corpus (the original manual and a supplementary manual for the tagged version). Printed manuals for the Brown Corpus cannot be obtained from Bergen. Some information on the London-Lund Corpus is distributed together with copies of the text and the KWIC concordance for the corpus. Users of the London-Lund material are also recommended to consult J. Svartvik (ed.). *The London-Lund Corpus: Description and Research*, Lund University Press, 1990.

A manual for the Kolhapur Corpus can be ordered from: S.V. Shastri, Department of English, Shivaji University, Vidyanagar, Kolhapur–416006, India. The price of this manual is US $15 (including airmail charges). Payment should be sent along with the order by cheque or international postal order drawn in favour of The Registrar, Shivaji University, Kolhapur.

Conditions on the use of ICAME corpus material

The primary purposes of the International Computer Archive of Modern English (ICAME) are:

(a) collecting and distributing information on (i) English language material available for computer processing; and (ii) linguistic research completed or in progress on this material;

(b) compiling an archive of corpora to be located at the University of Bergen, from where copies of the material can be obtained at cost.

The following conditions govern the use of corpus material distributed through ICAME:

1. No copies of corpora, or parts of corpora, are to be distributed under any circumstances without the written permission of ICAME.
2. Print-outs of corpora, or parts thereof, are to be used for bona fide research of a non-profit nature. Holders of copies of corpora may not reproduce any texts, or parts of texts, for any purpose other than scholarly research without getting the written permission of the individual copyright holders, as listed in the manual or record sheet accompanying the corpus in question. (For material where there is no known copyright holder, the person(s) who originally prepared the material in

computerized form will be regarded as the copyright holder(s).)

3. Commercial publishers and other non-academic organizations wishing to make use of part or all of a corpus or a print-out thereof must obtain permission from all the individual copyright holders involved.
4. The person(s) who originally prepared the material in computerized form must be acknowledged in every subsequent use of it.

Editorial note

The Editor is grateful for any information or documentation which is relevant to the field of concern of ICAME. Write to: Stig Johansson, Department of British and American Studies, University of Oslo, P.O. Box 1003, Blindern, N-0315 Oslo 3, Norway.

ICAME Journal is published by the Norwegian Computing Centre
for the Humanities (NAVFs edb-senter for humanistisk forskning)
Address: Harald Hårfagres gate 31, P.O.Box 53, Universitetet,
N–5027 Bergen, Norway
Telephone: Nat. 05 212954, Int. + 47 5 212954

ISSN 0801–5775