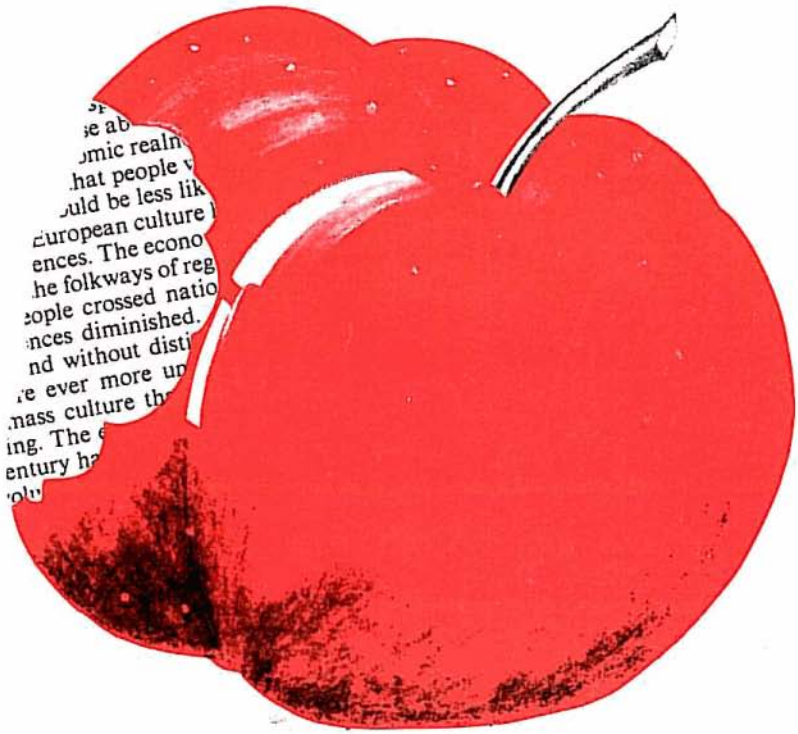
ICAME Journal

International Computer Archive of Modern English



Norwegian Computing Centre
for the Humanities

No. 11
April 1987

CONTENTS

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

Editor's Foreword

The change of name in the present issue from *ICAME News* to the *ICAME Journal* is a reflection of the gradual change of our publication, both in appearance and content. As in previous issues, we bring both articles and news items relating to computing in English language research. The articles by P.H. Peters and by K. Ahmad and G. Corbett report on corpora of Australian English. M.L. Owen takes up the problem of the evaluation of automatic grammatical tagging. R. Jones introduces a software package for the analysis of machine-readable texts. P. Collins presents results from his studies of cleft sentences in the LOB Corpus and the London-Lund Corpus. In this way, the present issue mirrors the various stages in computer corpus research – from the initial preparation of data, through the analysing stage, and to the finished product.

New developments within ICAME include preparations for the provision of material in new form (diskette, CD-ROM), the cooperation with the Oxford Text Archive and the Humanities Research Center at Brigham Young University, and the setting up of a program library and an electronic information service. See the "materials" sections and the report by Knut Hofland. The report from last year's conference in Amsterdam gives a picture of the varied concerns of researchers in computer corpus work.

So far, the operation of ICAME has been very informal and has primarily depended on the initiative of individual corpus workers and the support of the Norwegian Computing Centre for the Humanities. In spite of the apparent success of this informal modus operandi, we have decided to move in the direction of greater formalisation. Last year's conference saw the birth of an Advisory Board for ICAME. The following have agreed to be on the Board:

Jan Aarts, University of Nijmegen
Sidney Greenbaum, University College London
Jostein Hauge and Knut Hofland, Norwegian Computing Centre for the Humanities
Ossi Ihalainen and Matti Rissanen, University of Helsinki
Randall Jones, Brigham Young University
Henry Kučera, Brown University
Geoffrey Leech, University of Lancaster
Willem Meijs, University of Amsterdam
John Sinclair, University of Birmingham
Jan Svartvik, University of Lund

Starting with this issue, we will also – for the first time – ask for a subscription fee (see the enclosed leaflet). I hope that the new Board and the financial support through subscriptions will ensure the continued operation of ICAME, in the interest of researchers concerned with the computational study of the English language.

Stig Johansson
University of Oslo

Cleft and pseudo-cleft constructions in English spoken and written discourse

Peter C. Collins
University of New South Wales

This paper reports findings from a study based on the London-Lund (henceforth 'LL') and LOB corpora. One aim was to demonstrate, through an investigation of clefts and pseudo-clefts in natural discourse, the importance of taking into account the communicative properties of these constructions in analysing their structure. A second aim was to explore relationships between the communicative functions of the constructions and their distribution across the genres of LL and LOB.

The selection of LL and LOB as a database enabled broad comparisons of speech and writing to be made. It must be acknowledged, however, that the omission of handwritten and typewritten texts from LOB prevents that corpus from offering an entirely adequate representation of the written language.

Syntactic, semantic, and communicative properties

The constructions under review are exemplified in (1). Corresponding to the 'simple' or 'non-cleft' sentence (1a), there are in English sentences of the type (1b) (the 'pseudo-cleft') and (1c) (the 'cleft').

(1) a. Tom offered Sue a sherry.
b. What Tom offered Sue was a sherry.
c. It was a sherry that Tom offered Sue.

(1b) and (1c) are identifying constructions, in which the single clause of (1a) is divided into two parts: the identifier (which I shall refer to as the 'highlighted element'), namely *a sherry* in both cases; and the identified constituent (which I shall refer to as the relative clause), namely *what/that Tom offered Sue.*

Whereas many linguists (e.g. Prince 1978, Higgins 1979) restrict the class of pseudo-clefts to those like (1b) with fused relative clause introduced by *what* as subject, I have argued (Collins 1985) that those with relative clause introduced by the other *wh*-items of English, and those with relative clause introduced by *the* in conjunction with their pro-form equivalents, should also be included in the class (e.g. *The thing Tom offered Sue was a sherry; The one who offered Sue a sherry was Tom; That is why Tom offered Sue a sherry*).

As identifying constructions, (1b) and (1c) carry an implicature of exclusiveness not present in (1a) (which envisages the possibility that Tom might have offered Sue other things as well). In addition, (1b) and (1c) carry an existential

presupposition ('there is something that Tom offered Sue') that is lacking in (1a).

The primary function of 'pseudo-clefts' and 'clefts', as the names suggest, is thematic: they enable subsets of elements to be grouped into two parts in an almost unlimited number of ways (cp. *What Tom did was offer Sue a sherry; The one who offered Sue a sherry was Tom; It was Tom who offered Sue a sherry*). Furthermore the highlighted element and relative clause of pseudo-clefts may be inverted (e.g. *Tom was the one who offered Sue a sherry*). These constructions (pseudo-clefts with highlighted element as theme) are referred to as 'reversed pseudo-clefts' in the present study, while their non-reversed counterparts are termed 'basic pseudo-clefts'. There is a difference in the way that pseudo-clefts and clefts generate thematic prominence. In the case of pseudo-clefts the prominence is experiential, deriving from the structural equation of the highlighted element and relative clause. In clefts the prominence is textual, deriving from the structural device of predication, which introduces the theme as complement to the non-referential subject *it*. Clefts are not reversible, as are pseudo-clefts, so that the emphasis falls less upon the identity between the two parts than upon the predication of one part in the structure (the theme).

There are informational, as well as thematic, differences between pseudo-clefts and clefts. Basic pseudo-clefts display a consistently close mapping of the functions of theme, givenness and presupposition onto the subject relative clause, which is presented to the addressee as representing information that s/he should be prepared to accept as non-controversially recoverable. Basic pseudo-clefts thus offer the speaker a means of specifying precisely, before the announcement of the 'message', the background knowledge to which the addressee is expected to have access. The source for this knowledge may be cotextual or contextual.[1]

Cotextual recoverability may be quite straightforward (as in (2), where the only element of newness in the theme is the indication of affirmation in *really*: all else is directly retrievable from prior linguistic context), or it may be less direct, requiring the addressee to cooperatively infer an 'antecedent' (as in (3), where the reader is expected to infer that Mr Dixon's offering of something to his public follows from his role as a critic).

(2) #-the only thing I ever VÁRY#. you CÁN vary# is really [?] well you can vary ÀNYTHING# but the only thing I'm [?] -- *the thing that you RÉALLY vary#is*[ðiˈ] - HÒPS# (LL S.1.7, 253-4)

(3) My old friend and colleague, Campbell Dixon, used to tell of a conversation he had with a New York film-critic, a lady, who heard with an air of shocked incredulity that *what he offered his public was his own private and unsupported opinions*. (LOB A17,150-1)

As Table 1 shows, direct recoverability is favoured in LL, inferrable recoverability in LOB (cp. findings reported in Prince 1981). It would be interesting to explore

whether these differences relate to a general trend for writers to impose greater cooperative demands on their addressees than speakers.

Contextual recoverability may be of three types, deriving from the three variables that determine register. 'Field-antecedents' are retrievable from the setting of relevant events and happenings within which the language is functioning, as in example (4) below. 'Tenor-antecedents' are those that focus on the speaker/writer's emotions, reactions, thoughts and so on, which are presented as a legitimate concern of the cooperative addressee, as in (5). 'Mode-antecedents' are those which focus on the mode of communication itself, and are thus intrinsically metalinguistic, as in (6).

(4) D [ðiː ə] - the CHARACTERISTIC of our house#. is COFFEE cups#

C [m].

D and *what HAPPENS is# that you may make coffee for SIX# -- and put the empties in the SINK#* (LL S.1.12, 444-6)

(5) What I found depressing was the insistence that all the many good things in the country were due only to "socialism" and the Party and would not otherwise exist, together with fantastic ignorance of the western world or refusal to believe what did not suit the theory. (LOB B21,114-7)

(6) What chiefly stands out in this lively work, I think, is an accusation that Milton himself had smuggled into a later edition of Eikon Basilike the prayer, derived from Sidney's Arcadia, for which he then so resoundingly denounced King Charles in Eikonoclastes. (LOB J61,50-4)

Table 1 Basic pseudo-clefts and givenness in LOB and LL

	Cotextual recoverability		Contextual recoverability		
	Direct	Inferrable	Field	Tenor	Mode
LOB	37	87	22	33	13
	124 (64.6%)		68 (35.4%)		
LL	79	59	62	46	26
	138 (50.7%)		134 (49.3%)		

Whereas in basic pseudo-pseudo clefts it is the highlighted element that conveys the

'news', this element in the reversed construction is typically (in 82.2% of cases in the combined corpus) represented by a demonstrative with extended-text referential function (and thus inherently given, unless contrastive). The typically text referential function of the highlighted element (see Table 2), and the exclusive equation of this element with the backgrounded material in the relative clause, give rise to a special 'internal-referencing' discourse-function. With their capacity to relate together an anaphorically-referred-to chunk of preceding text with information presented by the speaker as 'not-at-issue', reversed pseudo-clefts are particularly suited to marking the conclusion of stages in the schematic structure of discourses. Along with cliches, generalisations, explicit repetitions and various other informationally-low forms, reversed pseudo-clefts serve appropriately as endings, as Stubbs (1983:24) has suggested, because they "provide no new information which can serve as a resource for further talk". Notice how Speaker A in (7) uses a reversed pseudo-cleft to draw together and conclude his description of the operations of a pawnbroker, before proceeding to offer a hypothetical example:

(7) A we would SÀY that[?]# we would KNÒW you SÉE that#[ə:m]. how much they'd got in a.local authority LÒAN or SÒMETHING#and SÒ#. we then would go out and SÁY#we now think you should GÈT this out of the local authority LÒAN and #.cos. X Y and Z looks like a good BÙY#-ÒH YÈS#we DÒ that #.that's the JÔB really#. oh it's quite a RESPÒNSIBÌLITY#--I mean not a PÈRSONAL RESPONSIBÍLITY# but a CÒRPORATE responsibility#-.

a yes

A *thaĭ's how it ÓPERATES*#.SÒ that I mean#if you were a CLÌENT of#of the FÌRM#-[əm ?] (coughs) I would have a FÍLE# (LL S.2.2, 749)

Table 2 Reference of highlighted element in reversed pseudo-clefts in LOB + LL

		that	*this*	other	TOTAL
Cataphoric/exophoric		13	12	6	31
Anaphoric	Non-extended	26	3	56	85
	Extended	316	95	30	441
TOTAL		355	110	92	557

There are three major informational types of cleft construction, one unmarked ('Type1'), and two marked ('Type2' and 'Type3'). Frequencies for these are presented in Table 3 below. Unmarked clefts display a correlation between the structurally highlighted element and the locus of new information ('new' more often in the sense of 'newly identified' or 'contrastive', rather than 'fresh'). For example:

(8) #it's really was BE↘RYL that did it I THI↗NK# (LL S.1.5,401)

Beryl is here being selected from a set of possible candidates: 'Beryl, and no one else you might be thinking of'.

A characteristic of unmarked clefts is the very low communicative dynamism of information in the relative clause (*that did it* in (8) is comprised of items with clearly anaphoric function). Often in fact the relative clause is so evidently recoverable that it is ellipsed, as in (9) where *you want to convince* is directly retrievable:

(9) B(...) #{A^ND {THO^SE}} ^ENGLISH Indologists# that I have met are in that I've TA↘LKED TO↗#- [ʔə] are MO↗ST enthusiastic#

a yes it's not the Indologists you want to convince

B [ʔm] -

a it's the people with money

B *it's the people with the MO↘NEY#* (LL S.2.1,871)

Often the contrastive function of the highlighted element goes hand in hand with an explicit formulation of contrast, e.g.:

(10) A he said you're sure *it's Marks and SPA↘RKS you're going to WO↗RK for#*

a,A (--laugh)

A *it's not just M and S↘ {SO↗MEWHERE#}* #
(LL S.2.12,1004,1006)

In one type of marked cleft (Type2) the relative clause contains new information and the highlighted element is typically short, and anaphoric or deictic. The very

length of the relative clause in (11) disqualifies it as a possible unmarked cleft, in view of the unlikelihood that a constituent of such size could represent given information:

(11) It should be remembered that until the implementation of the Guillebaud Report, under which railway rates of pay were based on the principle of "comparability" with those of comparable employees in other employments, railwaymen had worked for considerably debased rates of pay, and *it was they who had been providing the subsidy necessary for the running of the railways which are necessary to the economy of the country.* (LOB B11,56-8)

A second type of marked cleft (Type3) is that with a 'semi-new' item – typically a circumstantial, or 'scene-setting' adjunct of time, place, or the like – as highlighted element. Type 3 clefts have, unlike the others, the potential for discourse-initial distribution. They are classed as marked because, even though the highlighted element is non-given, it is the relative clause that is mainly responsible for conveying the 'message'. Examples follow:

(12) A TRIBUTE TO HAROLD CLAY
It is with deep regret that we pay a last tribute to a great friend and colleague who has passed on. (LOB F16,203-4)

(13) It is not by capturing more territory that science fiction will improve itself, (LOB G36, 90-1)

Type3 clefts were (as Table 3 shows) considerably more common in LOB than LL, and were particularly favoured in formal, learned genres.

Table 3 Informational types of clefts in LOB and LL

	Unmarked	Marked	
	Type1	Type2	Type3
LOB	195 (34.5%)	188 (33.3%)	182 (32.2%)
LL	76 (40.6%)	72 (38.5%)	39 (20.9%)
TOTAL	271 (36.0%)	260 (34.6%)	221 (29.4%)

Syntactic properties: numerical findings

Numerical findings relating to the syntactic characteristics of pseudo-clefts and clefts were noted to be interpretable in terms of the communicative features of the constructions described above.

Pseudo-clefts and clefts vary significantly in the class and function of highlighted elements. The present study confirmed Prince's (1978:885) claim that "the only significant overlap concerns focused NP's". As Table 4 indicates, 33.2% of basic pseudo-clefts selected a NP as highlighted element, 50.4% of clefts, and 90.0% of reversed pseudo-clefts. Pseudo-clefts strongly favour nominal elements (in the case of the basic construction, both phrasal and clausal: only 4% of the highlighted finite clauses were adverbial, rather than content, clauses). The preference is very strong with reversed pseudo-clefts, where *this* and *that* are selected even when the highlighted element functions as adjunct, in preference to *here, there, then, in this way*, etc. The findings for pseudo-clefts are compatible with the claim made above that pseudo-clefts encode an experiential form of thematic highlighting, deriving from the (reversible) equation of two elements in a coding relationship. Some typical examples follow:

(14) a. Mr Gaitskell said that *what stopped the Russians in the last resort from aggressive nuclear war was the certainty that they would be annihilated.* (LOB A04,142-4)

b. #what he DOESN'T REALIZE#is that not EVERYBODY else#can work quite as hard as HE can# (LL S.1.5,207-9)

c. This is what the Minister proposes. (LOB B14,77)

By contrast, clefts highlight virtually any item which is able to be thematised in the corresponding non-cleft. As well as NP's (as in 15a below), PP's (15b), finite clauses (15c), adverbial phrases (15d), and zero (where the cleft thematises an indication of tense, modality, polarity, etc., with all items having a representational function appearing in the relative clause, as in (15e), are represented in significant numbers. The fact that there are, amongst these items, several (zero and certain PP's) which could only be thematic in a theme-predicated structure and not in a pseudo-cleft (because they cannot be substituted by a WH-item) reinforces the claim that clefts display a different form (textual) of thematic prominence. Some typical examples follow:

(15) a. It was Mrs Kennedy who drew the crowds, said police. (LOB A28,26)

b. #it was [ə:] {THRÔUGH} DÂVID that [ə:m] #- ˇINGRID met Don# (LL S.4.4,148-9)

c. #--it is because GÔD#--has made the [su:mpri] supreme ŠACRIFICE#on our BEHˇALF#-that we are ÂBLE# to ask HˇIM#-to help us FˇIGHT against# -- the wiles of the DˇEVIL# (LL S.12.1,684-90)

d. It will be very seldom that permanent good can be done in this field under six months. (LOB H08,50-1)

e. If so, it must be that their God was more powerful than the Kikuyu's Ngai. (LOB K29,39-40)

Table 4 Class of highlighted elements in LOB + LL

	Pseudo-clefts		Clefts
	Basic	Reversed	
NP	154 (33.2%)	563 (99.0%)	379 (50.4%)
Finite clause	208 (44.8%)		51 (6.8%)
Non-finite clause	97 (20.9%)	2 (0.4%)	6 (0.8%)
Zero			109 (14.5%)
PP	5 (1.1%)	2 (0.4%)	162 (21.5%)
Adjective phrase		1 (0.2%)	1 (0.1%)
Adverb phrase		1 (0.2%)	44 (5.9%)
TOTAL	464 (100%)	569 (100%)	752 (100%)

Table 5 reveals differences in the weightings of the three primary functions of highlighted elements (subject, object, and adjunct). The functions that are most often thematic in ordinary non-cleft declaratives (subject and adjunct) are favoured in clefts, suggesting that their typical function is to imbue an already thematic item with further prominence (through predication). With basic pseudo-clefts the popularity of highlighted objects is predictable from English word order, but the popularity of subjects (vis-à-vis adjuncts) reflects a preference for participant-related functions. With reversed pseudo-clefts the preference for highlighted objects and adjuncts over subjects is the reverse of the situation with clefts, suggesting that

reversed pseudo-clefts achieve thematic prominence by selecting functions that are most unlikely to be thematic in ordinary declaratives.

Table 5 Syntactic function of highlighted elements in LOB + LL

	Pseudo-clefts		Clefts
	Basic	Reversed	
Subject	152 (32.8%)	81 (14.2%)	288 (38.3%)
Object	173 (37.3%)	218 (38.3%)	54 (7.2%)
Adjunct	27 (5.8%)	196 (34.4%)	276 (36.7%)
Complement of preposition	27 (5.8%)	57 (10.0%)	24 (3.2%)
Complement of subject	6 (1.3%)	17 (3.0%)	1 (0.1%)
Complement of verb	79 (17.0%)		
Zero			109 (14.5%)
TOTAL	464 (100%)	569 (100%)	752 (100%)

Genre distribution

Pseudo-cleft and cleft constructions are not evenly distributed in speech and writing (see Table 6). Pseudo-clefts greatly outnumber clefts in speech (by a ratio of 623:187, or 3.3:1), while clefts outnumber pseudo-clefts in writing (by a ratio of 565:410, or 1.3:1). Furthermore reversed pseudo-clefts outnumber basic in both speech (by a ratio of 351:272, or 1.3:1) and writing (by a ratio of 218:192, or 1.1:1).

Table 6 Pseudo-clefts and clefts in LOB and LL[2]

	Pseudo-clefts			Clefts
	Basic	Reversed	Total	
LOB	192 (1/5208)	218 (1/4587)	410 (1/2439)	565 (1/1770)
LL	272 (1/1599)	351 (1/1239)	623 (1/698)	187 (1/2326)

In Table 7 below figures are presented based on major subgroupings of the two corpora. In LL 'private' categories are those not recorded before an audience or on radio, including conversations between intimates and disparates, both face-to-face and via telephone, and 'public' categories include radio debates, interviews, commentaries, and audio-conditioned orations. In LOB 'informative' categories embrace press, religious tracts, popular lore, biography, government documents, scientific writings, and so on, while 'imaginative' categories cover a range of fictional genres.

Table 7 Pseudo-clefts and clefts in subgroups of LOB and LL[3]

		Pseudo-clefts			Clefts
		Basic	Reversed	Total	
LOB	Informative	140 (.37)	109 (.29)	249 (.66)	402 (1.07)
	Imaginative	52 (.41)	109 (.86)	161 (1.27)	163 (1.29)
LL	Private	161 (2.87)	277 (4.94)	438 (7.82)	121 (2.16)
	Public	111 (3.58)	74 (2.38)	185 (5.96)	66 (2.12)

Pseudo-clefts are comparatively more popular than clefts in the private categories of LL, with the results for the former determined largely by the figures for the reversed construction. The favoured conversational mode for reversed pseudo-clefts is face-to-face, with intimates or equals as participants, as represented in Category B, where the text frequency was 5.7: their frequency was slightly lower – 5.4 – in Category A where some of the conversations include interlocutors of disparate status, and significantly lower – 2.4 – in Category C where the channel is telephone rather than face-to-face (for further details on these categories see Svartvik and Eeg-Olofsson 1982). Further evidence for the influence of tenor relationships on pseudo-cleft ratios is to be found in a comparison of the three text categories of telephone conversation. In S.7 (conversations between personal friends) the basic: reversed ratio was 4:9 (or 0.4:1), in S.8 (business associates) it was 17:12 (or 1.4:1), and in S.9 (disparates) it was 9:3 (or 3:1).

The ratio of pseudo-clefts to clefts decreases as we move from the private categories (where it is 438:121, or 3.6:1) to the public categories (185:66, or 2.8:1). The differences are more striking if we compare the dialogic categories (A-E) of LL (where the ratio is 570:154, or 3.7:1) to the monologic categories (F-H) of LL (where it is 53:33, or 1.6:1). It would seem that such factors as the formality of the

speech situation and the extent to which the speech is premeditated exert a stronger influence on the frequency of clefts than pseudo-clefts. Indeed the category which had the highest frequency of clefts in LL (2.9 per text) was H (prepared oration: dinner speeches, sermons, lectures, and court cases).

While the overall ratio of clefts to pseudo-clefts in LOB was 565:410, or 1.4:1 (see Table 6), in the informative categories clefts were comparatively even more popular, the ratio being 402:249, or 1.6:1 (see Table 7). By contrast, in the imaginative categories the frequency of pseudo-clefts rivals that of clefts. It is the popularity of the reversed construction in fiction that is largely responsible for the difference. Reversed pseudo-clefts were almost entirely restricted to passages of dialogue, with authors exploiting the fact that the construction is one that is favoured in informal speech (cp. the findings for LL), in their attempts to recreate the spoken language. Another possible explanation for the increase in popularity of pseudo-clefts (particularly reversed) from informative to imaginative prose, is that fiction exhibits a 'speech-like' degree of context dependency in the extensive reference that is made to aspects of the internal temporal and physical situation that its author constructs (cp. Rader 1982).

Within the informative genres, where clefts were almost twice as popular as pseudo-clefts, the frequency of clefts appeared to vary along a dimension that might be characterised as 'factual versus opinionative'. Consider the following ranking of Categories A-J, ordered according to the frequency of cleft constructions:

H	(Miscellaneous documents, reports, etc.)	.36
A	(Press:reportage)	.54
J	(Learned and scientific writings)	.73
E	(Skills, trades, and hobbies)	1.02
F	(Popular lore)	1.25
G	(Belles lettres, biography, essays)	1.35
D	(Religion)	1.35
B	(Press:editorial)	1.40
C	(Press:reviews)	1.52

A strong determining factor in the frequency of cleft constructions here appears to be the extent to which the function of a text is on the one hand primarily descriptive (the aim being to present factual material objectively), or on the other hand the extent to which it is 'opinionative' (the aim being to mount arguments, engage in persuasion, mix opinion with fact, and so on).

Conclusion

Let us posit several explanations for the differences in distribution between pseudo-clefts and clefts, in terms of the communicative differences between the constructions. The popularity of pseudo-clefts (especially reversed) in speech is attributable largely to their givenness-orientation. Basic pseudo-clefts, we have seen, attach special status to background material, presenting it in the form of a subordinate clause which not only embodies a presupposition at the logico-semantic level, but also represents the theme at the textual level. It is not unduly surprising that a construction which so explicitly represents the background knowledge which the addressee is expected to be aware of, should occur more frequently in speech than writing. The basic-pseudo-cleft functions then as an interpersonal 'tracking' device within the flow of discourse. In this connection it is interesting to note comments by some linguists (e.g. Higgins 1979) on the interrogatival properties of basic pseudo-clefts, and proposals by some (e.g. Nakada 1973) that the construction be derived from an underlying structure containing an embedded question.

The communicative properties of reversed pseudo-clefts explain their popularity in speech (particularly in informal conversation between friends). The internal-referencing function of the construction, along with its generally low informativity, are well suited to the dynamic organisation of spoken language. The typical realisation of theme as a text-referential demonstrative enables a stretch of prior discourse (whose extent is likely to be larger with interlocutors whose acquaintance enables them to share a pool of common knowledge) to be identified with low-communicatively-dynamic information in the rheme/relative clause. Furthermore the almost clicheic quality of many reversed pseudo-clefts, deriving from their minimal newsworthiness, is apposite in informal spoken genres.

Cleft constructions exhibit several properties that account for their popularity in written discourse (and particularly 'rhetorical' genres). By contrast with pseudo-clefts, their orientation is towards newness. New information is highlighted, via thematic predication, both in unmarked clefts and in marked Type2 clefts. Even though imbued with a non-controversial flavour, the new information in the relative clause of marked constructions is considerably higher in communicative dynamism than that of reversed pseudo-clefts. The denser information-packing of writing therefore provides one form of explanation for the generic distribution of clefts. A further possible explanation relates to the lack of stress-marking in writing. Clefts may be used by the writer as a means of directing the reader into a particular interpretation of the information structure (namely one where the locus of new information is mapped onto the theme). It is precisely in 'opinionative' texts that one would expect writers to find a need for linguistic means whereby intonation might be indicated, and in these that the contrastive implication generated by the theme/new combination would prove an attractive resourse. A third reason for the comparative popularity of clefts in writing may be their structural similarity to

impersonal constructions (such as 'It is said that...', 'It is true that...'), from which they derive a depersonalised quality and a formality that is often out of place in casual spoken genres.

Notes

1. As Table 1 indicates, whereas cotextual and contextual sources were almost identical in number in LL, 138 and 134 respectively, in LOB there were almost twice as many cotextual, the ratio being 124:68, or 1.8:1. These findings presumably reflect general mode differences between speech and writing.
2. Bracketed figures represent the frequency of constructions per number of words. The figures enable frequencies to be compared across the two corpora (LL comprises about 435,000 words, and LOB about 1,000,000 words).
3. For details of the private/public subclassification for LL, see Svartvik and Eeg-Olofsson (1982). Bracketed figures represent number of tokens per text. (In LL, texts are about 5,000 words in length; in LOB they are about 2,000 words.)

References

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

Higgins, F.R. 1979. *The pseudo-cleft construction in English.* New York: Garland Publishing, Inc.

Nakada, S. 1973. Pseudo-clefts: What are they? *Papers from the ninth regional meeting of the Chicago Linguistic Society.* Chicago: Chicago Linguistic Society, 428-41.

Prince, E.F. 1978. A comparison of *wh*-clefts and *it*-clefts in discourse. *Language* 54:883-906.

Prince, E.F. 1981. Toward a taxonomy of given-new information. In *Radical pragmatics,* ed by P. Cole. New York: Academic Press.

Rader, M. 1982. Context in written language: The case of imaginative fiction. In *Spoken and written language: exploring orality and literacy,* ed by D. Tannen. Norwood, N.J.: Ablex.

Stubbs, M. 1983. *Discourse analysis: The sociolinguistic analysis of natural language.* Oxford: Basil Blackwell.

Svartvik, J. and M. Eeg-Olofsson. 1982. Tagging the London-Lund corpus of spoken English. In *Computer corpora in English language research,* ed by S. Johansson. Bergen: Norwegian Computing Centre for the Humanities.

Evaluating automatic grammatical tagging of text

M. L. Owen
Acorn Computers, Cambridge

1. Introduction

This paper is intended as a contribution to the general issue of evaluating work being done on an Esprit Project[1] in text-to-speech and speech-to-text conversion. Here, we limit ourselves to the evaluation of systems for the automatic grammatical tagging (AGT) of corpora, and in the course of this, explore briefly what the tagging process involves. In particular, we shall consider the effect of the tagset used on the success rate achieved.

Existing AGT systems, notably those connected with the Brown Corpus (Francis 1980), the LOB Corpus (Leech *et al* 1983), and the London-Lund Corpus (Svartvik & Eeg-Olofsson 1982), express their success rate for output prior to manual post-editing quantitatively, in terms of a percentage of correctly assigned tags, where 'correct' should be read as 'matching the judgement of the linguist who devised the system of labels'. Thus, for example, Leech *et al* report a 96.7% success rate for Lancaster's suite of programs. At first sight these measures seem entirely unproblematic, but the purpose of this paper is to explore the interpretation(s) that can be placed on them.

We should begin by making certain crucial distinctions. First, there will be a group of linguists who have taken on the task of collecting the corpus, transcribing it if it is a spoken corpus, making it available in printed or machine-readable form, and if their resources permit and their aims extend that far, labelling the forms in the text with their grammatical category (tag), and assigning a constituent structure to the sentences of the text. We will call this group the corpus compilers. The evaluation that is significant for them is of the type described above, as reported for the Lancaster project: we describe this as 'internal' evaluation.

The second group of linguists is of course the corpus users. Suffice it to say at this point that these two groups may or may not be the same, though it is likely that certain individuals will belong to both groups. At some stage, however, there will be users who have had no contact with the compilers. The evaluation performed by this group will be predominantly qualitative, since a quantitative answer cannot be given to the question 'how useful is this labelled corpus for my research?'. We call this 'external' evaluation, and consider it first in what follows.

2. External evaluation

While it is probably the case that the majority of linguists do not use corpus-based data in their research, anyone who wants to make inductive generalisations about usage needs something of the sort on which to base their claims. For the sake of reliability a corpus needs to be large (of the order of a million words), and the larger the corpus, the more essential it becomes that there are at least some automated search procedures. Clearly as far as single word-forms are concerned, a relatively simple concordance will suffice, though the assignment of forms to lexemes – if that is what is required – is by no means straightforward (Matthews 1974:27-31). Above this level, however, a system of grammatical labelling is a minimal requirement, and ideally, a full constituent parse is what is needed.

Suppose, for example, that a linguist is interested in relative clauses in English. He might wish to know not only what the grammatically permissible constructions are – this much he can ascertain by traditional, intuition-based methods – but what the patterns of usage are in modern English. In this case he might wish to locate all relative clauses within a corpus. To the extent that relative clauses are introduced by relative pronouns, he can locate the constructions he wants by a search for the forms *who*, *which*, etc. Many instances will turn out to be interrogative pronouns, and it will be necessary to weed these out by hand: a much less laborious process, nevertheless, than a manual search of the entire corpus. A grammatically tagged corpus, however, in which relative pronouns were distinguished from interrogative, would render this stage unnecessary.

However, some relative clauses have no relative pronoun, and a word search will not locate them. Only if the corpus is parsed (ie with constituent structure assigned as well as grammatical tags, and constituents such as relative clauses identified) can a search of this kind be conducted. The corpus user will evaluate the success of the corpus in terms of his own requirements, assessing how easily it enables him to locate all the instances of the constructions in which he is interested.

Such external evaluation is something over which the corpus compiler has no direct control; he will use his professional imagination, of course, but the labelling/parsing he produces cannot be expected to meet the needs of everyone who might come to use the corpus in the future. It is, after all, possible to conceive of many types of non-surface phenomena that a linguist might wish to locate: particular illocutionary acts, for example. The transformational grammarian of the early 1960s might have wanted to find instances of sentences that had – putatively – undergone certain transformations. This possibility, at least, is now one that the corpus compiler need not contemplate. In sum, any linguistic analysis that stands at some remove from surface features will be assisted only to a limited extent by the use of a corpus; this observation is not intended critically but merely to illustrate how harsh external evaluation might, in principle, be.

These points were made, in essence, in my review of Svartvik & Quirk's printed edition of part of the London-Lund Corpus (Owen 1982). In practice, of course, the

corpus user is likely in most cases to be able to formulate his requirements in the terms used by the corpus compilers, and where he cannot - in pragmatics research, for example - he will probably have to assemble his own corpus anyway.

External evaluation, then, is straightforward, even though it is not made in terms of measurements. A tagged corpus is useful to the linguist to the extent that (a) he is asking the kind of questions for which a corpus is relevant, and (b) he can formulate his questions in an appropriate way.

3. Internal evaluation

To explore this notion it is helpful to outline the steps undertaken in the construction of an automatically tagged corpus. Since our own project does not call for the assignment of constituent bracketing, but only for the tagging of individual orthographic words, we shall limit discussion to these aspects. The process can be broken down into seven steps.

1. Collect a corpus and convert it into some machine-readable form.
2. Assemble the set of tags to be applied: this involves crucial decisions concerning the type and number of tags, since these cannot easily be changed later.
3. Tag by hand a portion of the corpus (the 'learning' corpus). This process draws on the linguist's specialised knowledge, for although the tagging process applies only to one word at a time – there is no phrasal level – he will use his knowledge of constituent structure in selecting between alternative tags for an item. To take a simple example: in the following two sentences the item *that* will have different tags, and its pronunciation will vary with the tag:

   a. I remembered that yesterday (deictic pronoun)
   b. I remembered that yesterday he arrived (subordinating conjunction)

   In (b), the vowel of *that* is normally reduced to schwa, whereas in (a) it is not. For both sentences, the linguist has to examine the structure of the sentence as a whole in order to select the correct tag. Information about adjacent words would only be relevant insofar as it revealed that structure, and in these examples, not only are the words on either side of *that* the same orthographic words, they would also receive the same tag; the linguist therefore must look beyond the immediate environment of the word he is labelling. No principled limit can be placed on how many words in either direction he must examine, since this is structurally-determined. We will call the knowledge the linguist uses in making these choices LK1: it is an explicit, linguist's, statement of the knowledge possessed by the non-linguist native speaker.
4. Assemble a set of statistical statements concerning the relative frequency of

   a. each tag
   b. each possible form/tag pairing, and

c. each possible transition between tags.

This set of statements constitutes, in a very unusual sense from a linguist's point of view, a body of linguistic knowledge. We will call it LK2; its exact form and content is a function of all the choices made in stages (1)-(3) above.

It is essential to appreciate that it is knowledge about a corpus, not knowledge about a language, and that it is a type of knowledge that cannot, in principle, be held by a human speaker of the language. We should not, however, rule out the possibility that the human language user makes use, in processing, of certain heuristics bearing some resemblance to LK2; some evidence that this takes place is reported by Warren and Marslen-Wilson (in press).

5. Taking the entire corpus (minus the learning corpus) as input, label each form with all its possible grammatical tags.[2]
6. Devise and implement an algorithm for selecting the most probable tag for each ambiguously tagged item; this involves choosing the best path through a series of alternative tags, and an algorithm for this process is fully described in Atwell (1983). It is of crucial importance to appreciate that steps (5) and (6) use only LK2; the machine has no access to LK1, though this is the knowledge from which LK2 is derived.
7. Evaluate the success of the automatic tagging procedure. At this stage, the linguist brings his structural knowledge LK1 into play again – recall that the automatic tagging procedure has not had access to this. Each potentially tag-ambiguous item will be examined, and the validity of the selected tag assessed. It is this procedure that results in the percentage success rates mentioned at the beginning of this paper, and it is because it does not relate to any externally-imposed criteria that we have termed it 'internal' evaluation.

4. Factors influencing internal success rates

4.1 Choice of grammatical tag set

The size of the tag set does not, *per se*, influence success one way or the other. More important is the extent of ambiguous form/tag assignments. Thus a very small tag set in which one commonly-occurring form could be assigned any one of the tags in the set would result in a higher level of ambiguity than a larger set with very little ambiguity.

Furthermore, there are complex relationships between the relative probabilities of a given tag

a. occurring in the corpus overall (eg 20% of all tags are instances of tag B)
b. being assigned to a given item (eg 25% of all occurrences of item k are labelled as B's)
c. being preceded by some other specific tag.

To illustrate this, let us consider the implications of alternative decisions regarding the tag(s) assigned to the form *to*. The Brown and LOB Corpus compilers distinguish between the infinitive particle *to* and the use of the form as a preposition, but it would be possible simply to assign them to one category, or even go a step further and adopt the strategy they use for many frequently-occurring forms, and create a category TO which would be unique to that form. It is difficult to see in advance what the costs and benefits of these alternatives would be: how hard is it likely to be for the system to decide in any particular instance, and will the sub-classification help in the tag-assignment of any frequently adjacent items? The following examples are designed to illustrate these problems, but are also selected because the decisions made will affect the pronunciation of one of the words in the sentence (*contract(s)*). The resolution of such ambiguities is one of the external criteria imposed on the tagging process within the Esprit project.

a. they agreed to the contract
b. they agreed to contracts that their predecessors had rejected
c. they agreed to contract
d. they agreed to contract after contract

A text-to-speech system, receiving these sentences as input, needs to know whether *contract* is a noun (N) or a verb (V) in order to assign stress, and hence vowel quality. In (a), the presence of the article will so strongly favour N that this example is not problematic. In (b), if the remainder of the sentence after *contracts* could be identified as a relative clause, this would similarly favour the assignment of N to *contracts*. In the absence of such evidence, however, it would be helpful to know whether *to* was a particle or a preposition: this would enable V to be assigned in (c), and N in (d). It is impossible to tell whether the presence of *after* following *contract* would assist in resolving the ambiguity; certainly as far as the human processor is involved, it is only on reading the remainder of the sentence that he may have to re-label the first occurrence of *contract* as N rather than V, but our system does not look this far ahead.

The sub-classification of *to* would therefore help in the resolution of certain tag-assignment ambiguities. However, on what basis can the system assign these more finely-differentiated tags to *to*? In examples (b)-(d), the following item does not help, since it is itself ambiguous (it is, after all, in order to resolve this very ambiguity that we are considering sub-classifying *to*); in (b) and (d) there is additional material in the sentence, but it is not necessarily the case that the system will be able to take advantage of this. All the forms occurring before *to* are unambiguously verbs, and there are sentences for which, if we break down this category to show the type of complementation each V can take – whether NP, VP, some other category, or more than one of these – we could resolve the ambiguous labelling of *to*. For example, in

e. they hoped to contract
f. they walked to market

if we knew that *hope* takes a VP complement, we would know that *to* was a particle, and thus that *contract* was a verb; similarly in (f), *walk* can take a PP complement, *to* is therefore a preposition, and *market* a noun. Verb sub-classification will thus resolve ambiguity in these examples, but only if we also sub-classify *to*, thus effectively using it to transmit information down the chain. (We assume a first-order process only.)

Returning to examples (b)-(d), subcategorisation of verbs would not at first sight appear to be useful, since *agree* can take a wide range of complements: VP (example (c)), PP (examples (a) and (b)), and even NP ('agree the figures'). Success here in transmitting the right information down the chain, via the tagging of *to*, as far as the subsequent N/V-ambiguous item, will depend solely on the relative frequency of occurrences of *agree* with each type of complement in the learning corpus.

To summarise: decisions concerning the size and composition of the tagset have consequences that can only be discovered through the lengthy process of hand-tagging the learning corpus and applying the resulting LK2 to a large test corpus. Tags that are chosen on the basis of their inherent linguistic interest – because they provide a vocabulary in which significant grammatical statements can be made – may also be the right tags to choose for the sake of a high internal success rate, but they may not.

4.2 Type of tags used

If any tags are defined in terms of the category they follow or precede, tag-assignment will be more successful. Suppose we decide that item b is tagged as follows:

J when preceding c;

K otherwise.

In this extreme case, success is almost, though not entirely, guaranteed.[3]

What this amounts to is the claim that certain aspects of LK1 are expressible in terms appropriate to LK2, though this is by no means the same as saying that they are the same rules, or even necessarily of the same form, for it is a property of natural languages that they are structure-dependent. However, some items may be tagged in ways not entirely unlike this. For example, the class of items occurring in determiner position includes true determiners (especially the articles), pre-determiners (*all, both, half*), and post-determiners (*other, few*). Some of these forms can appear in more than one category, for example

a. few children who came to the party were disappointed
b. the few children who came to the party were disappointed

In (a), *few* is a determiner, in (b) a post-determiner, on the simple and obvious grounds that it does indeed follow the determiner *the*. The matter is not as simple as this, of course, since *few* may also be tagged as a pronoun in sentences such as

c. many are called, but few are chosen
d. few of the children enjoyed the party

but in neither of these examples can *few* be a post-determiner.

Generally speaking, the more tags there are in the tagset whose application is determined (for the human labeller) by conditions stated in terms of *adjacent* items, the greater the success rate that will be achieved.

It should be noted that in constructing a grammar (ie a representation of LK1), linear ordering does not have the central role it has in LK2; indeed, arguments exist to show that constituency and linear ordering should be represented separately (Gazdar *et al* 1985: 44-50).

Another way of ensuring at least partial success is the strategy mentioned above in the discussion of *to*, that is, the selection of tags that are not grammatical categories at all, but merely labels applying unambiguously to a small class of forms, or even to a single form. The Brown corpus tagset, for example (Francis 1980) contains several such tags, including the following:

form	*tag*
be	BE
been	BEN
does	DOZ

No other forms have these tags assigned to them. The resolution of some potentially difficult choices in this Gordian way may affect the internal success rate substantially, (without necessarily reducing the usefulness of the tagged corpus to the corpus user), provided that differentiation of these items is not crucial in the transmission of information down the tag 'chain'.

4.3 Overall extent of ambiguity implicit in the tagset

Given a corpus and a tagset, and making the assumption that forms tagged unambiguously are correctly tagged, the proportion of words with unambiguous tag-assignment forms the baseline for internal evaluation: if, for example, 70% of words are unambiguously tagged, the internal success rate cannot be less than 70%, at least for words that occur in the learning corpus and are therefore in the dictionary. The position of this baseline will, furthermore, be determined by the number and type of tags used. An internal success rate of 95% is from the point of view of the ambiguity-resolving software less impressive if the baseline is 85% than if it is 60%, in that the system is resolving fewer ambiguities.

In addition, consider the following circumstances. Suppose that only 10% of words appearing in the dictionary built out of the learning corpus have more than one label. Suppose also that the frequencies with which different labels are applied to a given word tend to be strongly 'skewed' in favour of one label rather than another, then the system clearly has stronger information on which to base its decisions. For example, if in this situation each ambiguously-labelled word has only

two possible labels, and that in each case label A is applied nine times to every one application of B in the learning corpus, then simply by choosing the most frequent label on every occurrence of an ambiguous word, an internal success rate of 99% will be achieved.

5. Comparison of AGT systems using internal measures

The general lesson from these observations is simply that comparison of different AGT systems using internal measures is far from straightforward, since the success rate achieved depends on so many independent but interacting decisions.

It should also be noted that the probabilistic approach to tagging will work best of all with languages whose linear order is relatively fixed. To my knowledge, the only languages for which work of this type has been done are not at one extreme or the other of this spectrum, but this may nevertheless be another factor affecting internal success rates.

6. External evaluation revisited

We are now in a position to return to the notion of external evaluation introduced earlier. We defined this as evaluation imposed by a user standing at some considerable remove – in time, space, or research stance – from the corpus compilers. It is also likely, however, that the corpus compilers themselves had some application in mind for the tagged corpus, and to the extent that it serves their purposes it can be regarded as successful. This type of evaluation is still external, in that it cannot be measured quantitatively, but it lies within the control of the corpus compilers in a way that true external evaluation does not.

As an example, consider the requirement that for Esprit project 860, grammatical tagging should help in the disambiguation of homographs and homophones for grapheme-to-phoneme and phoneme-to-grapheme conversion respectively. AGT will then be successful to the extent that it resolves these ambiguities and makes the correct choices. However, an additional aim is an improvement in the prosody assigned to sentences in text-to-speech conversion, and an AGT system, however successful internally, is of limited value here, since without the next stage of constituent parsing information on phrasal boundaries is not available. Atwell (1983), however, describes how this may be attempted while still maintaining the probabilistic approach. This work is now in progress at the University of Leeds under the direction of Prof. G Sampson.

Notes

1. No 860: The Linguistic Analysis of Six European Languages
2. We do not discuss here the source of this information: a simple dictionary based on the learning corpus can be used, together with a morphological module proposing labels for words not occurring in the learning corpus.

3. It is conceivable that AGT could fail for the sequence *b c* just in case other factors contributing to the choice of J or K had a powerful countervailing effect. Recall that this is a protocol for the human labeller, not something accessible, as such to the AGT process. All that process knows about is what is in fact the case in the data derived from its learning corpus, not what is permissible.

References

Atwell, E. S. (1983): Constituent-likelihood grammar. *ICAME News* 7, 34-66.

Francis, W. N. (1980): A tagged corpus: Problems and prospects. In S. Greenbaum, G. Leech & J. Svartvik (eds), *Studies in English linguistics for Randolph Quirk*, London: Longman, 192-209.

Gazdar, G., Klein, E., Pullum, G. & Sag, I. (1985): *Generalised phrase structure grammar*. Blackwell.

Leech, G., Garside, R. & Atwell E. S. (1983): The automatic grammatical tagging of the LOB Corpus. *ICAME News* 7, 13-33.

Matthews, P. H. (1974): *Morphology*. Cambridge University Press.

Owen, M. (1982): Review of J. Svartvik & R. Quirk, (eds) (1980): *A corpus of English conversation. Journal of Linguistics* 18.2, 436-442.

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

Warren, P. & Marslen-Wilson, W. D. (In press): Continuous uptake of acoustic cues in spoken word recognition. To appear in *Perception and Psychophysics.*

Towards a corpus of Australian English

P.H. Peters
Macquarie University

Although the differentness of Australian English from American and British English has long been recognised, it is only relatively recently that studies of Australian English have been based on sizable bodies of data. Those which have been – most notably the Mitchell-Delbridge survey of adolescent speech published in 1965, and Horvath's study of the sociolects of Sydney (1985) – have been concerned with phonology. Such studies of Australian grammar and morphology as there have been, have based themselves on rather limited data. Differences in the Australian lexicon have been more comprehensively documented since Morris' *Dictionary of Austral English* (1898), and again by Baker in *The Australian Language* (1945), and most recently in the *Macquarie Dictionary* (1981). Yet all these have been based on lexicographical citations rather than any broad data base.

The lack of a comprehensive data base has spurred linguists at two universities in Sydney (P.C. Collins at the University of New South Wales, and D. Blair and P.H. Peters at Macquarie University) to set up a corpus of contemporary Australian English, as a resource for lexical, grammatical and discourse information. There is some incentive to structure the basic corpus along similar lines to the Brown and LOB corpora (matching them for size and spread of samples), so as to permit direct comparisons between the three varieties of English. Admittedly, there may be a geographical as well as a temporal dimension of difference, since both the existing corpora were based on texts of 1961, and the Australian one will consist of texts from 1986. But the source texts are in all cases written ones, and changes in the written language are less rapid than those of the spoken. We therefore expect the regional differences to be more significant than any temporal ones.

A sample corpus like the Brown or LOB is of course relatively small if one is interested in lower frequency items of vocabulary or grammar. (See Francis 1982:11-14.) It is however adequate for the more superordinate words of the lexicon (or "semantic primitives" to use Dik's term, (1978)), and for the function words. Such a corpus can offer insights into some of the more frequent grammatical structures, and by relative frequencies, show up the preferred ways of realising particular grammatical notions in a particular dialect or style.

The 2000 word samples of the Brown and LOB corpora are rather arbitrary units for researching the larger structures of discourse. The need for monitor corpora comprising much longer texts is now well recognised (Sinclair, 1982), and in compiling the Australian corpus we intend to put down material for the sample and the monitor corpus simultaneously – ie. extracting a 2000 word sample from any

suitable text received in machine-readable form, and also depositing the whole text in our archives.

1. Deposition of texts

In most western countries, the publishing industry is turning to computer typesetting, with increasing numbers of texts available in machine-readable form. We had hoped to take advantage of this trend and to obtain all our raw material direct from magnetic tapes and disks, saving time and error in the process. Unfortunately, not all Australian publishers yet use computer typesetting. The government publishers generally do, but some of the key commercial publishers of textbooks and schoolbooks (books with complicated formats) are not yet able to arrange the material satisfactorily by computer. Not all the daily newspapers yet use computer typesetting either.

Unless we were prepared to bias our sampling in favour of publishers who used computer typesetting, we had to face the fact that our data gathering would have to be by a combination of methods. Computerised sources might be used when available and convenient for our purposes (see next section on the latter point), but some keying in of texts was unavoidable. Optical scanning might occasionally relieve the typist, although reports on the use of the Kurzweil Data Entry Machine (KDEM) by Renouf (1985:10) show that it has limitations in scanning newspaper text (where pilot work has shown we particularly need help), and in reading some very glossy publications. Optical scanners may improve in the course of time, but more sophisticated machines are unlikely to cost less than the KDEM. The challenge for us then will be in securing access to that expensive machinery – too pricy for our project alone to purchase.

If we were willing to postpone the project for say ten years, the Australian publishing industry might be more fully computerised, and the corpus builder could probably count on getting all texts in machine-readable form. As things are, we are obliged to take samples from both computerised and non-computerised sources.

2. Pilot work on newspaper text

To develop basic computational techniques for the Australian corpus, we have been experimenting with a large volume of computerised newspaper text (approximately a quarter of a million words) supplied on magnetic tape by two metropolitan newspapers. For reasons which will be explained below, the tapes are not a suitable source from which to obtain 2000 word samples in set subcategories. They have nevertheless been invaluable as experimental raw material, and we are indebited to the *Sydney Morning Herald* and the *Adelaide Advertiser* for them.

The news text was received in the form of "dump" tapes – the form in which it could most easily be supplied, but which requires extensive manual editing before any samples can be systematically extracted. In dump tapes the articles are not put down in any particular order, or even grouped according to subject, and one

concerned with sport from the last page of the paper may be juxtaposed to a political item from, say, page 5. In fact we could only discover the page on which an article appeared by reference to the printed copy. The page location of every article had to be checked and noted, and every article had to be read, to decide its classification in terms of the Brown/LOB subcategories – political, financial etc. Reading and checking showed that a few of the articles on the Sydney tape were a little longer or shorter than the version which appeared in print; and on the Adelaide tape, several versions of the same article were stored, so that the editing process there involved deleting all but the final version. (Some of the pre-final versions would have appeared in earlier editions of the day's paper, but it seemed best to work with the copy from a final edition of the paper.)

So in neither case did the tapes comprise a straightforward representation of the articles (all and only) of a day's newspaper. For the newspaper house they are no doubt a flexible resource against which to keep the presses rolling, but for us with the sample corpus in mind, they are less than a convenient source. Only by extensive pre-editing (and by reference to the printed copy) could we use such tapes to systematically call up, say, all the political articles on page 5, in order to take a composite 2000 word sample from them. The ease of acquiring newspaper text in this form seems to be outweighed by the difficulty of manipulating it, if one intends to sample it in the same way as the compilers of the Brown and LOB corpora did.

The experience of working with the dump tapes convinced us that when seeking newspaper extracts (categories A - C) for the sample corpus, it would be simpler to work straight from the printed copy. With the various other categories (D - R) sampled by Brown and LOB, machine-readable texts are probably more accessible and tractable. Being more permanent publications, the final form is put down sequentially on the publisher's tape, and the extracting of a 2000 word sample is a simpler matter.

Apart from their experimental value, the dump tapes also have a value as source material for the Australian monitor corpus. The fourth section of this paper presents some of the results obtained so far in lexical comparisons between the Sydney and Adelaide papers.

3. Sampling of Australian newspapers

Alongside the computational work on the *Sydney Morning Herald* and the *Adelaide Advertiser*, questions as to how to sample Australian newspapers more comprehensively remained with us, and how best to parallel the sampling of Brown and LOB in the Australian context. The Australian newspaper market at present resembles that of the USA rather more than the British, in the absence (or rather) shortage of national newspapers. The metropolitan newspapers are the dominant publications, and there are only two national daily newspapers: i) *The Australian*, and ii) *The Australian Financial Review*, whose circulations are in each case much smaller than that of any of the major metropolitan dailies. (See Table 1 below.) There are suburban and country regional newspapers in each state, but they are

given over to a handful of local events and masses of advertising. (The suburban ones are therefore often handed out free of charge.) They are of little importance as sources of written news language, and we felt it reasonable to leave them out of our sampling, and to concentrate on the metropolitan presses, and such national press as there is.

Among the metropolitan dailies there are both highbrow and lowbrow publications. The latter are marked with asterisks in Table 1, and because they often have big circulations, it seems important to cover them. In terms of any "reception index" (Francis 1982:10) they are at least equal to their highbrow counterparts in the eastern states of Australia. It also seems important to cover the dailies in every state, including the less populated states such as Tasmania and the Northern Territory, because within their capital cities the local daily wields influence, even if its overall circulation is small by comparison with dailies in other states.

In our sampling of Australian newspapers we are therefore departing from a purely random selection of the total newspaper inventory - though in different ways from those adopted by Brown and LOB. (Brown's sampling was based on those kept on microfilm in the New York Public Library (Francis 1982:17), while LOB gave extra weight to the national as opposed to the provincial press (Johansson, Leech & Goodluck 1978:14,16).) Our weighting is designed to ensure that the major daily in all capital cities is sampled, and that in cities where there is more than one daily, all are sampled in proportion to their 1986 Audit Bureau of Circulation figures.

Table 1. Proposed sampling of Australian daily newspapers for category A.

Daily newspaper	Circulation 1986	No. of Samples for category A
National		
The Australian	134,000	1
The Australian Financial Review	66,000	1
New South Wales		
Daily Mirror*	296,000	3
Daily Telegraph	265,000	2
The Sun*	258,000	2
Sydney Morning Herald	255,000	2
A.C.T.		
Canberra Times	45,000	1
Victoria		
The Age	233,000	2
The Herald	237,000	2
Sun News-Pictorial*	549,000	5
Queensland		
Courier-Mail	217,000	2
Daily Sun*	133,000	1
Telegraph*	119,000	1
South Australia		
Adelaide Advertiser	211,000	2
The News*	159,000	1
West Australia		
The West Australian	238,000	2
Daily News*	98,000	1
Tasmania		
The Mercury	55,000	1
Northern Territory		
Northern Territory News	18,000	1
Total:	3,586,000	33

The same principles of sampling (representing all states, and in proportion to circulation figures) will be used with Sunday and weekly papers, and in seeking samples for categories B and C.

4. Lexical trends in the *Sydney Morning Herald* and the *Adelaide Advertiser*

One of the claims made from time to time about Australian English is that it is becoming americanised, though the signs of it are not particularly conspicuous, and no full-scale study of it has yet been mounted (Leitner 1985:56). With our deposits of newspaper text from Sydney and Adelaide, we have a quite sizable lexical sample against which to make comparisons with the lexis of the Brown and LOB corpora:

Sydney Morning Herald	(17/10/85)	97,187	running words
Adelaide Advertiser	(24/9/85)	67,824	" "
Adelaide Advertiser	(27/9/85)	67,248	" "
	Total	232,259	" "

Both papers are heterogeneous in content, and the items are written up by many different journalists, as evidenced by the frequency of bylined articles. There is room for individual stylistic and linguistic choice within the general register of newspaper journalism. The orthographic form of words may be constrained by editorial or in-house style (see below), but the actual choice of words is relatively unconstrained – whatever terms the writer deems fit for communicating with the local community. So newspaper writing enshrines the comings and goings of many words in current usage. Arguably, it is a more up-to-date index of linguistic trends than most other kinds of writing.

Some of the word frequencies in the Sydney and Adelaide papers make very interesting comparisons with their counterparts in the Brown and LOB corpora, particularly those which appear to be indexical of American/British differences from Hofland and Johansson's analysis (1982:33-38 and section 8). Table 2 presents in columns 1 and 2 the raw frequencies of selected words from Brown and LOB which also occurred quite often in the *Sydney Morning Herald (SMH)* and *Adelaide Advertiser (AA)*. The total occurrences of each word in the two Australian newspapers appear in the fourth column. The third column contains a pro rata equivalent to the raw frequency (adjusted in terms of occurrences per million words), to permit ready comparison with the Brown and LOB frequencies.

Table 2. Comparison of selected words in the Brown and LOB corpora, and in a corpus of two Australian newspapers.

word	Brown	LOB	Aust. (adj. per 1,000,000)	Aust. (raw frequency)
new	1635	1181	1806	419
old	660	670	366	85
last	676	870	1866	433
first	1361	1287	1401	325
city	393	199	806	187
town	212	262	190	44
village	72	126	47	11
committee	168	230	362	84
council	103	343	1362	316
shall	267	348	17	4
will	2244	2269	4302	998
should	888	1276	948	220
would	2715	2682	3551	824
maybe	134	82	69	16
perhaps	307	406	52	12
holiday	17	74	56	13
vacation	47	1	4	1
film	96	162	185	43
movie	29	4	82	19
shop	63	84	91	21
store	74	42	130	30
journey	28	63	4	1
trip	81	37	112	26

The Australian frequencies often seem to polarise the relationship shown in the Brown and LOB data, intensifying the differences in one or both. Occasionally there is an apparent reversal of a joint American/British trend, though this may be a matter of generic limitation (the fact that the Australian frequencies are drawn only

from newspaper sources), rather than a regional difference. A case of polarisation may be seen in the Australian frequencies for *new* and *old*, and the Australian frequencies for *last* and *first* are an instance of reversal. However, both *new* and *last* are more frequent in categories A to C of LOB than in the overall British corpus. Compare the theoretical frequency of *new* in A - C: 1952 with the overall raw frequency of 1181; and *last* with (theoretically) 1703 in A - C, but overall (actually) 870. (The theoretical figures are from Hofland and Johansson (1982) section 7.) Both words rank more highly in the Brown category A data analysed by Zettersten (1978) than they do in the overall corpus. In A *new* is rank 37 and *last* rank 46, whereas they rank at 64 and 137 respectively in the whole Brown corpus. It is not hard to see that their frequencies in newsprint are likely to be inflated by the press's concern anywhere in the world with the immediate past and with new developments.

The same generic explanation could account for Australian polarisation of the relationship between *city, town,* and *village* (very strongly in the direction of the first) and between *committee* and *council* (in favour of the latter). City events, and the decisions of urban councils, are the everyday stuff of metropolitan papers. But we might also note that the American/British differences with *council* as opposed to *committee,* and with *city, town,* and *village,* were both statistically significant (at .001 level) and symptomatic of regional variation. The Australian preference for *council* and *city* in those two groups endorses neither the American nor the British exactly, and it is in such independent selections from the available variants that Australian English asserts its own dialectal identity. The integration of American with British linguistic habits has certainly emerged at other points in this analysis of Australian newspaper text, as we shall see. They create a distinctive local mix which may well prove to be the hallmark of Australian English, once the details have been tested on a multi-generic corpus.

The Australian frequencies of *shall/will* and *should/would* once again seem to polarise the trends shown in Brown and LOB, but most strikingly with the first pair. *Shall* appears to be obsolescent in Australian news language, whereas it has (or had?) a definite role to play in the Brown and LOB texts. It is true that *shall* occurred more often in the legal, scientific and religious texts of Brown and LOB than in the other categories of informative prose (Krogvig and Johansson 1984:76), and also notably in British romantic fiction; and the Australian counterparts of these kinds of writing have yet to be analysed. Still the frequency of *shall* in Australian printed news is substantially below its theoretical equivalent in LOB categories A - C (17 versus 152 per million), and even in the category A material from Brown, its theoretical frequency was 52 per million (Zettersten, 1978). The loss of *shall* and the intensified use of *would* (against *should*) both suggest that Australian practice is more in line with American trends.

One further pair of function words on which Americans and Britons appear to differ is *maybe/perhaps*. *Maybe* is much more in evidence in Brown, although even there it is considerably outnumbered by *perhaps*. The Australian data so far shows a

striking reversal of the preference for *perhaps* shown by both LOB and Brown, and *maybe* outnumbers *perhaps*.

When we turn from such function words to specific content words, we again find examples in which Australian modes of expression seem to be developing along American lines. The development is less evident with *holiday/vacation* (just a slight turn of the tide), but quite clear on *film/movie*, where the latter is on the increase. The trend with *shop/store* is definitely in line with Brown rather than LOB, and on *journey/trip* the Australian frequencies intensify the differences shown in Brown, and imply only minimal use of *journey*.

So far we have spoken of Australian English as if it were a homogeneous entity. The newspaper raw material actually allows us to make som tentative interstate comparisons, and to see that the preference for an American variant may be stronger in one state than the other. For instance, while the overall Australian figures for *shop/store* show that the latter is gaining ground, the trend is much stronger in Adelaide (or at least in the *Adelaide Advertiser*) than in the Sydney newspaper. Their respective values are:

	SMH	*AA*	Difference coefficient
shop	11	10	+0.05
store	3	27	-0.80

A chi-square test shows that the differences are indeed significant: $X^2 = 10.6$, df. = 1, p = <.01.

Adelaide also seems to take the lead in the use of *maybe* rather than *perhaps*:

	SMH	*AA*	Difference coefficient
maybe	4	12	-0.50
perhaps	8	4	+0.33

and once again the chi-square test confirms that the differences are statistically significant: $X^2 = 5.38$, p = <.05.

By contrast, it is the Sydney paper which leads the trend with *film/movie*:

	SMH	*AA*	Difference coefficient
film	25	18	+0.16
movie	15	4	+0.58

But by chi-square test, these differences are not yet highly significant: $X^2 = 2.96$, p = <.10 only. One would hesitate to suggest that American influence is stronger in Adelaide than Sydney, though different details of American usage may well be in vogue in different states, as these examples seem to show.

The same conclusion emerges if we examine some of the orthographic practices of the Sydney and Adelaide papers. The *Adelaide Advertiser* uses spellings such as *color/favoritism/flavored* when the *Sydney Morning Herald* would use *-our* in these and all their counterparts. The *SMH* meanwhile makes a practice of spelling words

such as *sizable* without a medial *e*, when the *AA* uses *likeable, sizeable* etc. These differences certainly represent editorial decisions rather than any consensus among journalists in either state (Peters 1985:38-40). But again, it is interesting that Australian newspaper editors make different selections from the possible American spelling variants. Neither paper, it should be noted, is American-owned (at least when this paper was being written), so these spellings cannot be explained in terms of corporate management. (Cf. Leitner's comments (1984:56) on the American spellings in the Murdoch-owned newspapers in Australia.) We might also note that when the *SMH* and the *AA* converge on a particular spelling variant, it is sometimes the British and sometimes the American one. Both make consistent use of *-ise* in verbs such as *recognise* and *victimise*, and use *centre, defence,* and *offence,* all of which go against American practice. Yet both endorse American practice with *judgment* and *program.*

In matters of morphology, the two Australian newspapers once again use a mixed bag of British and American variants, as the following examples from the Australian corpus show:

Table 3. Some morphological selections made by Australian newspapers. (The raw frequencies for each are contained in brackets.)

i)	burned (1)	burnt (6)	ii)	-	afterwards (4)
	dreamed (1)	-		-	backwards (3)
	-	leant (1)		downward (1)	-
	-	leapt (3)		forward (28)	-
	-	learnt (8)		-	onwards (1)
	spelled (2)	-		outward (1)	-
	spoiled (1)	-		toward (4)	towards (24)
				upward (1)	upwards (7)
iii)	among (50)	amongst (1)	iv)	older (9)	elder -
	while (166)	whilst (1)		oldest (5)	eldest -

The numbers of instances are mostly small, and differing grammatical roles may contribute to the overall variation in the first two paradigms, eg. adjective v. adverb with *upward/upwards*. Yet the overwhelming preference for *forward* (presumably as both adjective and adverb), makes a noticeable contrast with the several others for which *-wards* is preferred. The mixed morphological practices of parts i) and ii) of the table are on balance more British than American. In contrast, the preferences of parts iii) and iv) are indisputably with American usage.

5. Interim conclusions

The data discussed shows a number of points at which Australian English (as evidenced in newspaper writing) is moving away from the norms of British English (LOB) and developing in line with those of American English (Brown). The americanisation is however intermittent, and inconsistent in some of the sets and paradigms we have looked at. One could perhaps argue that this very variability in the selection of linguistic variants, and the differences from one state to another, are a sign of the pervasiveness of American English (via mass media) in Australia. Its resources provide a reservoir of alternative expressions for Australians to tap, and to make their own selections among. On the other hand, the patchiness of the American selections shows that there is no wholesale capitulation to the American dialect. The phonology of Australian English is certainly not americanised, and with only intermittent use of American lexical and morphological variants, it still seems an overstatement to say that Australian English is becoming americanised. Rather, selected American variants are being assimulated here and there into the fabric of Australian English, and losing their American flavour in the process.

The signs of differing lexical frequencies in different states need further investigation. It will be interesting to look for them in the Melbourne-Surrey corpus (see Ahmad and Corbett, in this issue), as a sizable volume of Victorian journalism – albeit consisting entirely of editorials.

All the points raised here emerged from newstext, and from just two sources of it. With samples from a comprehensive range of Australian newspapers (as described above in section 3), we may hope to confirm the trends observed so far, or see them balanced out. Beyond that, our task is to sample Australian writing in the numerous other corpus categories. Newspaper texts may nevertheless, after all that, prove to be the most up-to-date indicators of changes in Australian usage.

References

Ahmad, K. and G. Corbett. 1987. The Melbourne-Surrey Corpus. *ICAME Journal* (this issue).

Baker, S.J. 1945. *The Australian language.* Revised 1966. Sydney: Currawong Press.

Dik, S.C. 1987. *Stepwise lexical decomposition.* Lisse: Peter de Ridder Press.

Francis, W.N. 1982. Problems of assembling and computerising large corpora. In Johansson S. (ed.) 1982:7-24.

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

Horvath, B. 1985. *Variation in Australian English.* Cambridge: Cambridge University Press.

Johansson, S. (ed.). 1982. *Computer corpora in English language research.* Bergen:

Norwegian Computing Centre for the Humanities.
Johansson, S. with G. Leech and H. Goodluck. 1978. *Manual of information to accompany the Lancaster-Oslo/Bergen corpus of British English.* Oslo: University of Oslo, Department of English.
Krogvig, I. and S. Johansson. 1984. *Shall* and *will* in British and American English: A frequency study. *Studia Linguistica* 38:70-87.
Leitner, G. 1985. Australian English or English in Australia – Linguistic identity or dependence in broadcast language, *English World-Wide* 5:55-83.
Macquarie Dictionary. 1981. Ed. A. Delbridge. Sydney: Macquarie Library.
Mitchell, A. and A. Delbridge. 1965. *The speech of Australian adolescents.* Sydney: Angus and Robertson.
Morris, E.E. 1898. *Dictionary of Austral English*. Facsimile 1972. Sydney: Sydney University Press.
Peters, P.H. 1985. A survey of spelling variants in Australian newspaper style guides. *Working Papers of the Speech, Hearing and Language Research Centre, Macquarie University*: 35-52.
Renouf, A. 1984. Corpus development at Birmingham University. In *Corpus linguistics*, ed. J. Aarts and W. Meijs, Amsterdam: Rodopi, 3-39.
Sinclair, J.M. 1982. Reflections on computer corpora in English language research. In S. Johansson (ed.) 1982: 1-6.
Zettersten, A. 1978. *A word-frequency list based on American English press reportage*. Copenhagen: Akademisk Forlag.

The Melbourne-Surrey Corpus[1]

Khurshid Ahmad and Greville Corbett
University of Surrey

Introduction

The Melbourne-Surrey Corpus is a corpus of newspaper texts of Australian English, now available through ICAME. In this paper we give a brief description of the corpus, followed by an account of its use to date.

The corpus

The corpus consists of texts taken from The Age, a quality daily newspaper published in Melbourne. We are grateful to Mr Peter McLaughlin, Editorial Manager of The Age, for permission to make the corpus available for research purposes. The texts are all editorials which appeared from September 1, 1980 to January 30, 1981. This was an interesting period, including an election among other events, and the editorials range over a wide variety of subjects. Within the dates given, the editorials were taken on a random 93 days, to give a total of 100,000 words of text. The original motivation for assembling a corpus with these specifications was to allow easy comparison with results obtained by Nixon (1972); he used a manual corpus of 100,000 words of editorials taken from The Times to investigate corporate concord – the agreements found with nouns like *committee (the committee has/have decided)*. Plural agreement in such instances is markedly less common in Australian English than in British English.

The editorials for each date selected (there are typically two per day) are stored together in a separate file (thus there are 93 such files). There is a separate word count of each file, which is available from ICAME with the magnetic tape.[2] The material is all in upper case, which has some advantages for scanning. If upper/lower case information is required for particular items, this can be got from the copies of the originals, also lodged with ICAME. The computer texts retain the line breaks of the originals, so that matching to the originals takes a minimum of time. It is hoped that this corpus will be of value to those working on the varieties of English, and in particular will complement the work being done on spoken Australian English (Clark & Fraser 1982).

Use of the corpus

The main use of the corpus to date has been in work which lies on the border between language teaching and linguistics. We have been involved in computer

assisted language learning (CALL) since 1976, more recently with various collaborators, including Margaret Rogers (Surrey) and Roland Sussex (Melbourne). We wanted to give advanced learners a means to explore foreign languages. Advanced learners know that some rules are variable (corporate concord in English, mentioned above, is one such rule). However, to find sufficient examples to give an insight into the regularities involved would take a great deal of time. For this reason a package called SEARCHSTRING was developed. It allows a student sitting at a terminal to produce a concordance interactively. This package does not, of course, offer all the facilities of concordance programs such as CLOC (Reed 1978) or the Oxford Concondance Program (Hockey & Marriott 1980); that is not its purpose. Rather it provides the student with examples in three lines of context, together with the source, and it does so rapidly. Once ten examples have been found, the student is given the option of continuing or of finishing the search. Besides the texts discussed here, the student can scan Russian or German texts. For more details of SEARCHSTRING, see Ahmad, Corbett & Rogers (1985); some preliminary thoughts on computer corpora for teaching purposes were recorded in Ahmad, Corbett, Rogers & Sussex (1985:126-7). A sample run of SEARCHSTRING is given in the appendix.

Conclusion

We hope that this small addition to the material held by ICAME will be of value to researchers, and that some readers will be promted to consider the potential of corpora for teaching as well as for research purposes.

Notes

1. Corbett wishes to thank the Council of the University of Melbourne for the award of a Research Fellowship (1980-81), and the Economic and Social Research Council (UK) for a Personal Research Grant (1985-86) number C00232218. Neither award was primarily for preparing the Melbourne-Surrey Corpus; in fact it occupied a small proportion of the research time. However, without the first award the project would not have been undertaken, and but for the second it would not yet have been completed. The last stages of preparation were facilitated by a grant from the Centre for Translation and Language Studies, University of Surrey, which is also gratefully acknowledged.
2. Besides the number of words, the number of lines, the number of characters, and the maximum line length are given for each file, as well as for the corpus as a whole. We are grateful to John Hodder for use of his wordcounting program.

References

Ahmad, K., G.G. Corbett, & M. Rogers, 1985. Using computers with advanced language learners: An example. *The Language Teacher* (Tokyo) 9, 3:4-7.

Ahmad, K., G.G. Corbett, M. Rogers, & R.D. Sussex. 1985. *Computers, Language*

Learning and Language Teaching. Cambridge University Press.
Clark, J.E. & H. Fraser (eds.) 1982. *Australian Speech Archive* (Speech and Language Centre, Occasional Papers). Macquarie University.
Hockey, S. & I. Marriott. 1980. *Oxford Concordance Program, Version 1.0, User's manual*. Oxford University Computing Service.
Nixon G. 1972. Corporate-concord phenomena in English. *Studia Neophilologica* 44:120-6.
Reed, A. 1978. *CLOC user guide* (BUG 13). Birmingham University Computer Centre.

Appendix

Demonstration run of SEARCHSTRING

OK, RUN SEARCHSTRING [typed by student]

This program is designed to enable you to search through a large amount of text to find examples of particular words in context. Material is available in the following languages:

(1) English
(2) Russian
(3) German
Type 1 for English
or type 2 for Russian
or type 3 for German

1 [typed by student]

Do you want to:
(1) see the results of the search on the screen as the program runs?
or
(2) wait to look at the results until the search is over?
Type 1 or 2

1 [typed by student]

Do you want a print-out of the results of the search? Type YES or NO

NO [typed by student]

Input character string

GOVERNMENT [typed by student]

The string you asked for is:
GOVERNMENT

Providing this is the string you want, press RETURN. However, if you

have made a typing mistake, type ERROR and press return, in order to correct it.

[student presses return]

Do you want the search to include inflected forms? Type YES or NO.

NO [typed by student]

Scanning ENGLISH corpus for GOVERNMENT

- Variety :AUSTRALIAN -
- Text type :NEWSPAPER -
- Subject area :POLITICS -
- Filename :AGE1 -

Current file contains no examples of "GOVERNMENT"
>>>>>> Press RETURN to scan the next file <<<

Scanning ENGLISH corpus for GOVERNMENT

- Variety :AUSTRALIAN -
- Text type :NEWSPAPER -
- Subject area :POLITICS -
- Filename :AGE2 -

WASHINGTON PRESSED FOR THIS FACILITY OR WHETHER
THE FRASER GOVERNMENT LOBBIED FOR THE REQUEST.
DESPITE SOME FEARS THAT THE PRESENCE OF

OF SHARED INTERESTS. BUT, AS IN THE 1950S AND 60S,
THE PRESENT COALITION GOVERNMENT SEEMS
OVER-EAGER TO LOCK THIS COUNTRY INTO THE ROLE OF A

GREATER INDEPENDENCE AND SELF-RELIANCE IN ITS
DEFENCE PREPARATIONS, AND THE GOVERNMENT SHOULD
EXPLAIN HOW ITS LATEST DECISIONS FIT IN WITH THIS

Scanning ENGLISH corpus for GOVERNMENT

- Variety :AUSTRALIAN -
- Text type :NEWSPAPER -
- Subject area :POLITICS -
- Filename :AGE3 -

OF PLACE OF TRUST EITHER CIVIL OR MILITARY" IN VICTORIA TODAY. THE STATE GOVERNMENT FEARS IT DOES. IF SO, AS THE PREMIER, MR HAMER, HAS

PROHIBITION, IF IT STILL BINDS THE STATE, ARE QUITE INTOLERABLE. THE STATE GOVERNMENT HAS HAD ITS SUSPICIONS FOR SOME TIME, AND LAST YEAR ASKED

SUSPICIONS FOR SOME TIME, AND LAST YEAR ASKED THE FEDERAL GOVERNMENT TO LEGISLATE TO ALLOW THE STATES TO REPEAL OBSOLETE IMPERIAL LAWS

IS THAT THE STATUTE OF WESTMINSTER OF 1931, WHICH ENABLES THE COMMONWEALTH GOVERNMENT TO REPEAL REPUGNANT BRITISH LEGISLATION, DOES NOT

AFFAIRS. DESPITE ITS APPREHENSION, THE VICTORIAN GOVERNMENT HAS ONLY RECENTLY DECIDED IT SHOULD ACT OVER THE ACT OF SETTLEMENT. THE CHAIRMAN

OR HER OTHER REALMS.
THE STATE GOVERNMENT MUST URGENTLY ANSWER THE QUESTION WHETHER THE ACT OF

IMMEDIATE ACCESS TO THE REGISTER. OTHERS, INCLUDING GOVERNMENT BACKBENCHERS AND OPPOSITION MPS, WILL HAVE TO OBTAIN MINISTERIAL

>>>>>>> Press RETURN to continue <<<<

10 examples of the string "GOVERNMENT" have now been found. Do you wish to continue with the search? Type YES (to continue) or NO (to stop)

[at this point the student can continue, select a new string, or finish using the package]

Accessing the Brown Corpus using an IBM PC

Randall L. Jones
Brigham Young University

Even though the Brown Corpus is now more than 25 years old, it is still a very useful tool for the study of the English language. The rich variety of styles and topics contained in the 500 samples represents a diversity that reflects well modern American printed English. The principal creators of the corpus, Nelson Francis and Henry Kučera, were very careful to insure an accurate text which was coded with sufficient information.

The main problem with using the Brown Corpus has been getting information from it. There exists no standard retrieval program for use with the corpus, and even though it is now possible to obtain a concordance of the texts (e.g. the ICAME Brown Concordance), each user has specific interests which may or may not be served by a single concordance. The Brown Corpus was developed at a time when the word "computer" meant only mainframe computer. The expense of processing such a large body of data on a mainframe computer restricts greatly its practicality for many potential users.

With recent developments in microcomputers many of the restrictions associated with the processing of large texts no longer exist. Processing speeds of 8-12 megahertz, internal memory of 640 KB and above, and hard disk capacity of 20 megabytes and more are becoming common. It is more likely that a language specialist today will have access to such a powerful microcomputer than was the case with a mainframe ten years ago.

In this article I will report on a project in which the entire Brown Corpus has been converted from mainframe to DOS ASCII format, and then prepared for use with a PC text retrieval program. With the text retrieval program it is possible to locate occurrences of single words, lists of related words, phrases, substrings, and contextually defined groups of words in a matter of seconds. The accessed information can then be printed out directly or stored for further processing.

The transfer of data from mainframe to PC format was greatly facilitated by the use of DCA Corporation's Irma card and associated software. With the Irma card my IBM PC/XT can simulate an IBM 3279 terminal connected to our IBM 370/138 mainframe, where the ICAME Version 2 of the Brown Corpus is stored. The corpus was first divided into 2,000 line segments, then transferred to the XT using the Irma software. Translation from EBCDIC to DOS ASCII is automatic.

Because the 2,000 line segments did not necessarily correspond with the division of the 500 individual texts, it was necessary to edit each segment and restore the original boundaries. At the same time, the reference codes (e.g. A01 0001) were replaced by new codes consistent with the text retrieval program. The program

requires three levels of definition in order that the location of the accessed information can be referenced. For the Brown Corpus we chose to designate the highest level as the corpus itself, the next level as each of the 500 selections, and the third level as each line within the selection. Thus when an item is found it can be seen immediately that it is located e.g. in selection D7, line 134.

Aside from the adding of reference codes no additional processing of the text is necessary for indexing. We used a simple SNOBOL program to reformat the existing reference information in the corpus to be consistent with the indexing software. It is actually possible to omit the codes altogether, but their presence does facilitate the identification of the source of any located information.

The text retrieval program is called WordCruncher (formerly BYU Concordance) and was written by Monte Shelley and James Rosenvall of Brigham Young University as an interactive concordance program. It consists of two parts: IndexETC and ViewETC. (ETC refers to the company which now markets the software, Electronic Text Corporation.) IndexETC prepares a text by creating a kind of "road map" to the information in it. The user has the option under IndexETC of specifying a stopword list, i.e. words that are not to be concorded, e.g. *a, and, the* etc. It is also possible to change the sorting order of the character set, e.g. for use with foreign languages. The indexing procedure is a batch process and takes about 20 minutes per 100 KB of data.

Because the indexing takes place within memory there is a limit to the size of the text that can be indexed at any one time. This limit depends not only on the size of the text per se, but also on the number of unique words within the text. For example, a German document requires more memory than an English one of the same size because it will generate more unique words. After several parts of a large corpus have been separately indexed they can be merged together as one large indexed file, thus making it possible to search for information in the entire corpus at one time.

ViewETC allows the user to locate data in an indexed text. For example, within the Brown Corpus we can search for the following kinds of information: a single word, a list of words, a phrase, a substring, two or more words within a user defined context, or two or more lists of words within a user defined context. A few examples will illustrate how this is done.

When the ViewETC program is started the user is first asked to select from the so-called "Bookshelf" which text is to be accessed. In this case Brown is selected. One is then given the option of looking at part of the text, generating a printed index or concordance or searching for information in the text. If the third option is selected, i.e. searching for data, the program then displays a window in which eight of the headwords from the Brown Corpus are displayed together with the corresponding frequencies. It is possible to move around in this list by pressing the up and down arrows (one word at a time) the PgUp and PgDn keys (one window at a time), the Home and End keys (move to the top or Bottom of the list), or by typing in a word. Even if the list contains tens of thousands of words, as is the case

with the Brown Corpus, the access to any word is virtually instantaneous.

Let us assume that I am interested in seeing all occurrences of the word *proclaim.* After typing it in I press the RETURN key and immediately I see displayed on the screen the first six occurrences of the word, each within a three-line context. (The actual size of the display window can be defined by the user to be from one to 23 lines.) Above each window is the reference information explaining where each word is located in the text. By pressing the RETURN key once again the context for a single occurrence fills the screen.

It is likely that I am not only interested in the word *proclaim* but also the inflected forms *proclaims, proclaimed, proclaiming* and perhaps even *proclamation.* Instead of looking up each word separately I can create a list with all of them, and then view them together as a group. The order of the words as they are displayed reflects the order in which they occur in the text.

It may be the case that I wish to find a phrase consisting of several words, e.g. *on the other hand.* I can simply find all occurrences of any of the individual words, then select out only those that are part of the phrase. But there is an easier way. After I type in the word *on* I press the space bar, then type in *other*, again followed by the space bar, etc. until all words in the phrase have been entered. Once again the RETURN key is pressed and all occurrences of the selected phrase are displayed. This process requires a little more time, as the program must search for the intersection of each adjacent pair of words.

By pressing the asterisk key ("*") a substring can be defined, e.g. the suffix *-tion.* I can choose to find any word that contains this suffix either as the last characters (terminate with a period) or anywhere in the word (terminate with another "*"). After I have defined the substring the program begins looking in the headword list for the first word containing *-tion.* When it is located I can either view it in context, or else place it in a special list for later viewing.

Frequently it is desirable to find a word or group of words as they occur in the context of other words. For example, I might be interested in the verb *reported* followed by the conjunction *that.* Because other words can intervene between the two, it is not possible to foresee what the many possibilities might be, e.g. "She reported yesterday that the committee has decided...", "The mayor reported to the assembled group that it would be necessary to ...", "He reported several weeks after the article had appeared in the newspapers that the money had been taken by a member of the staff." ViewETC allows the user to enter two words or list of words, then define the context in which they should occur, i.e. List A and List B, List A but only if words from List B are not in the same context, List A or List B; List A only before List B, List A only after List B, List A before or after List B; within n characters, within the same line, within the same page.

After designated words have been located in the text the user has several options as to what to do with the data. It is possible, for example, to print out the words and context in the same format that is displayed on the screen. One can also mark a specific section of the text to be printed out. Instead of printing out the data it is

also possible to save it to a DOS file, then call it into a word processing program for later massaging. With certain kinds of software (e.g. WordPerfect Library) it is even possible to have ViewETC and a word processing language co-resident in memory and switch back and forth between the two. In such a case the data can be placed on a temporary "clipboard" and moved immediately from the ViewETC program to the word processor.

There are numerous other features of WordCruncher, but space limitations do not permit a complete description of the program. It works very well with the Brown Corpus, providing the user with virtually immediate access to any data as long as it can be defined lexically. Unfortunately, grammatical information is not easy to locate, as it is difficult to describe grammatical constructions on the basis of words. To look for past tense verb forms for example would require a listing of the individual verbs (*went, sang, did*, etc.). To look for *-ly* adverbs on the other hand would be relatively simple. (In the tagged version of the Brown Corpus one could, of course, simply look for the codes designating the desired grammatical features.)

It is exciting to see how rapidly language information such as that contained in the Brown Corpus is becoming accessible to so many users. Several times in the past few months I have received requests from colleagues around the world to provide them with examples from a variety of corpora. In most cases the search takes no more than a few minutes. The data can be printed out, transferred to a diskette and sent in the mail, or even transmitted on BITNET. Data bases such as the Brown Corpus provide us with a rich variety of samples for our use in linguistic and literary research. Programs such as WordCruncher can assist us in gaining rapid access to this information.

Note

Electronic Text Corporation is located on 5600 North University Avenue, Provo, Utah 84604, USA

ICAME 7th

The 7th International Conference on English Language Research on Computerized Corpora in Amsterdam, 9-11 June, 1986

The conference attracted over 40 participants from 8 countries: England, Holland, Finland, Norway, Sweden, Canada, Israel and China. As at previous conferences, there were papers reflecting the various stages of computer corpus work, from the compilation of corpora, through the analysis of machine-readable texts and the development of analytical tools, to the use of machine-readable texts for studies of particular aspects of the English language. See the list of papers below. For some of the papers we also give abstracts, as submitted by the authors. The proceedings of the conference have been published in Willem Meijs (ed.) *Corpus linguistics and beyond,* Amsterdam: Rodopi. Price: Dfl. 80.

Apart from the papers, there was a panel discussion on "Corpus linguistics in relation to other areas of research and application", with invited contributions from: S.C. Dik, Amsterdam, on "Functional grammar"; T. van Dijk, Amsterdam, on "Discourse analysis"; H. Kerkman, Nijmegen, on "Experimental psycholinguistics"; and L. Pols, Amsterdam, on "Speech technology". An evening session, chaired by Gert van der Steen, was set aside for a discussion of exchange of software. The result was that the Norwegian Computing Centre for the Humanities takes on the responsibility for the organisation of the exchange of software as part of its functions within ICAME. See Knut Hofland's report later in this issue.

There was a demonstration session, which included the use of Gert van der Steen's program QUERY and the LDB developed at Nijmegen. Both allow sophisticated searches in analysed text. Other programs presented were a search program for LDOCE and a program for phonetic transcription of speech. At the conference there was also established an Advisory Board for ICAME; see the Editor's Foreword.

The participants are indebted to Willem Meijs and the other members of the organising committee for a successful and well-organised conference. The next ICAME conference will take place in Helsinki in May 1987.

List of papers

Jan Aarts and Nelleke Oostdijk (Nijmegen) "Grammars in corpus analysis"
Karin Aijmer (Lund) "*Oh* and *ah* in English conversation" – see abstract
Eric Akkerman (Amsterdam) "ASCOT progress report" – see abstract
Bengt Altenberg (Lund) "Predicting text segmentation into tone units" – see abstract
Nancy Belmore (Montreal) "A pilot study of the application of corpus linguistics to the specification of word classes for language understanding systems" – see abstract

Ted Briscoe (Lancaster) "The use of the LOB Corpus in the development of a CPSG grammar of English"

Lou Burnard (Oxford) "CAFS: A text segmentation machine and some applications"

Jeremy Clear (Birmingham) "Monitor corpora"

Nina Devons (Jerusalem) "Observations on *one's* in contemporary American English" – see abstract

Pieter de Haan (Nijmegen) "Exploring the LDB: Noun phrase complexity and language variation" – see abstract

Hans van Halteren (Nijmegen) "Using an analysed corpus as a linguistic database"

Theo van den Heuvel (Nijmegen) "Interaction in syntactic corpus analysis" – see abstract

Ossi Ihalainen, Merja Kytö, and Matti Rissanen (Helsinki) "The Helsinki Corpus of English Texts" – see abstract

Stig Johansson (Oslo) and Knut Hofland (Bergen) "The tagged LOB Corpus: Description and analyses" – see abstract

Jan de Jong and Pieter Masereeuw (Amsterdam) "A new implementation of the LSP grammar"

Göran Kjellmer (Gothenburg) "Aspects of English collocations" – see abstract

Gerry Knowles (Lancaster) "Tone-grouping by numbers"

Antoinette Renouf (Birmingham) "Lexical resolution" – see abstract

Geoffrey Sampson (Leeds) "Evidence from the LOB Corpus against the grammatical/ungrammatical distinction" – see abstract

John Sinclair (Birmingham) "Upward and downward collocation"

Anna-Brita Stenström (Lund) "Carry-on signals in English conversation" - see abstract

Jan Svartvik (Lund) "Taking a new look at word class tags" – see abstract

Yang Huizong (Shanghai) "Automatic term identification"

Abstracts

Oh and *ah* in English conversation

Karin Aijmer
Lund University

In almost any conversation between two or more participants speakers use a lot of *oh* and to a lesser extent *ah*. These words are of special interest because they give access to the mental processes going on in the speaker's mind at the time they are taking place or immediately afterwards. The general meaning of *oh* is to express the speaker's reaction to an unexpected stimulus. We must however constrain this general description of its meaning in order to answer the following questions. Where in the conversation can it be inserted and why does it occur there? Is it equally frequent in all genres of (spoken) English? What is the difference in function between an utterance with *oh* and the same utterance with *ah*? The aim of my talk is to try to answer these questions by studying the occurrences of *oh* and *ah* in 34 informal conversations from the London-Lund corpus of spoken British English. The investigation shows that the distribution of *oh/ah* is dependent on properties of the preceding context. An explanation of why *oh* and *ah* are used is attempted in terms of the notion relevance. The hearer formulates his reactions to the preceding utterance so that he shows that he has adopted the relevant interpretation. It must also be possible for the speaker to indicate explicitly whether a prior utterance is relevant. This is done by means of *oh/ah*. The use of *oh/ah* marking relevance is dependent on the type of information given in a prior turn, whether it contains information which is open to dispute, or assumptions which go against the speaker's own assumptions. In the answer to a question *oh/ah* indicates that the expectations or predictions raised by the question are not fulfilled. The difference between *oh* and *ah* can best be shown by their different collocations. *Oh* occurs with collocations expressing a positive or negative evaluation; *ah* only with components expressing a favourable evaluation.

ASCOT progress report

Eric Akkerman
University of Amsterdam

ASCOT, which stands for Automatic Scanning System for Corpus Oriented Tasks (1), aims at the construction of a lexical data-base system and an associated scanning system, to be employed in (semi-)automatic syntactic analysis. Most of the information going into the system will be extracted from the computer-tape version of the Longman Dictionary of Contemporary English (henceforth LDOCE). In my

paper I reported on the work done in the second year of the ASCOT project, with an emphasis on the linguistic aspects of our activities (for our previous work, see Akkerman et al 1985). Computational work involved the development of a program with which the necessary information could be extracted from LDOCE and restructured into an optimally accessible format. Recently this program was finished and we are now in the process of creating the actual ASCOT lexicon and connecting it to the morphological analysis program Reroute.

Linguistic work was mainly concerned with the actual contents and form of the ASCOT entries. As far as the contents are concerned, we are of course very dependent on the information that is available in LDOCE. Therefore the grammatical coding system of LDOCE was analysed in considerable detail and often the grammatical approach of LDOCE was compared with that of Quirk et al in *A Comprehensive Grammar of the English Language* (1985). As a result of our critical assessment, a number of code combinations emerged that we found grammatically rather questionable or even incorrect. Therefore, for each code combination a decision was made how to treat it in the ASCOT lexicon. There were always three options:

i) the code could be adopted without change;
ii) the code could be adopted, but with minor changes and/or a certain warning note;
iii) the code could be rejected and any word with that code would be recoded for the ASCOT lexicon, if possible automatically.

It is impossible to discuss all our findings and conclusions here; for a detailed account, see Akkerman et al (forthcoming).

The form of the ASCOT codes is of course closely related to their contents. In general the codes consist of different information positions. They are built up hierarchically, so as to make the various types of information optimally accessible. The structure of the noun codes may serve as an example:

lemma	* word class	* type	* number	* countability + complementation	* position	* used with	* rest info

In LDOCE, the noun sugar has the following entry:

sugar /.../ n 1 [U] (...) 2 [C] (...) 3 [N] (...)

In ASCOT the entry will be:

sugar * N * C * var * U ; C * N * - * - *

indicating that it is a common (C) noun (N) which is variable for number (var), has both a countable (C) and an uncountable (U) meaning, and can also be used in the

vocative (N). Actual word forms (as they are found in an uncoded text) will be provided with a correct code by the Reroute program.

The ASCOT software package (containing a scanning system, the lexicon, and the morphological analysis program) will be available for bona fide research purposes in the course of 1987.

Note

(1) The ASCOT project is funded by the Dutch Organization for Pure Academic Research (ZWO) under project number 300-169-004.

References

Akkerman, E., Masereeuw, P.C., Meijs, W.J., 1985, Designing a computerized lexicon for linguistic purposes, *ASCOT Report* No. 1, Amsterdam, Rodopi.

Akkerman, E., Meijs, W.J., Voogt-van Zutphen, H.J., forthcoming (1987), *ASCOT Report* No. 2.

Predicting text segmentation into tone units

Bengt Altenberg
Lund University

An important task for the TESS project at Lund University is to establish a set of rules than can "chunk" a written input text into appropriate information units that will serve as the domain of subsequent rules of intonation assignment. This chunking process must simulate natural speech segmentation as far as possible, but it must be fully automatic and rely on a combination of punctuation cues, statistical probabilities and grammatical information produced by the parser. These conditions are not easily reconcilable, but using a predictive model of speech segmentation (adapted from Crystal 1975:15-22), it is possible to reduce the principles of tone-unit segmentation in a prepared and partly scripted monologue from the London-Lund Corpus of Spoken English to a set of grammatically defined rules that show promise of working fairly satisfactorily. These rules operate cyclically in top-down fashion, assigning tone-unit boundaries between clauses, clause elements and phrase constituents.

The coverage and success rate of these rules vary from one type of structural boundary to another, but truly unpredictable cases are comparatively rare (in the monologue examined so far), and there is good hope that other cases of insufficient coverage or rule failure can be reduced or eliminated when the rule system is enriched by segmentation data from a larger corpus.

The rules are derived entirely from spoken data and thus do not make use of punctuation cues in the input text, but it is obvious that punctuation, however unreliable it may be as a guide to segmentation, will provide an additional aid in

certain cases.

One problematic feature of the rule system is its top-down procedure. Apart from theoretical difficulties of such an approach (e.g. to what extent it reflects natural speech segmentation even when a text is read aloud), the demands it makes on accurate high-level grammatical parsing (distinguishing for example different types of clause boundaries) are obvious. It is possible, however, that this top-down approach can be replaced by a "shallower" procedure, but this alternative has not been investigated.

Reference

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

A pilot study of the application of corpus linguistics to the specification of word classes for language understanding systems

Nancy F. Belmore
Concordia University, Montreal

With the introduction of the QUERY program for pattern extraction, Meijs, van der Steen and their colleagues at the University of Amsterdam made it possible to considerably enhance the areas of application of corpus-based linguistic research. This paper describes a pilot study in one such area, viz. in the development of informationally-relevant part-of-speech categories for language understanding systems. The particular problem selected was the classification of words ending in *-ed* when they occur as pre- or post-noun modifiers. From an informational standpoint, such words are an interesting group because even their assignment to the major categories of verb and adjective can be a vexed question. Thus, the Brown and LOB tagging systems, while largely compatible, differ markedly in the tagging of many *-ed* words when they occur in such patterns.

In this pilot study QUERY was used to extract from the tagged Brown corpus *-ed* words tagged JJ (adjective) or VBN (past participle) which occur as pre- and post-noun modifiers. It was then used to extract from Brown instances of the major sentence types from which pre-noun *-ed* modification structures can be derived. The *-ed* pre-noun modifiers which had been extracted were then manually sub-classified in terms of their relation to these sentence types. This resulted in twelve preliminary *-ed* sub-classes. The small number of *-ed* post-noun modifiers extracted from the corpus did not warrant their sub-classification in such terms.

Methodologically, the study has shown the necessity to at least partially automate any manual sub-classifications of QUERY outputs so as to achieve both consistency

and rigor. It also suggests that the informationally-relevant categorizations required for language understanding systems will almost certainly necessitate numerous sub-categorizations of the traditional major parts of speech and the definition of new major categories.

Observations on ONE'S in contemporary American English.

<An analysis of occurrences in the Brown Corpus.>

Nina Devons
Hebrew University of Jerusalem

The FREQSUCON (Devons 1985) entry ONE, following the regular pattern, consists of the headword and its variant and compound forms: *one, ones, one's, no one, no one's, anyone/'s, everyone/'s, someone/'s, someone'll* and *oneself.* Three main sense varieties are discriminated: i) numerical, ii) = a person, iii) replacive (anaphoric or ostensive) (Devons 1986). All occurrences of the variant ONE'S fall within subentry ii. This paper addresses questions concerning the use/interpretation of ONE'S (+ ONESELF) which do not apply to the other lexical forms of the subentry, including discussion of what appear to be points of divergence between AmE and BrE usage.

Of the 65 occurrences listed in Kučera and Francis (1967), 62 are instances of the possessive form of the indefinite pronoun (human reference) ONE (as distinguished from *no one's*, 2 occ., and *the great one's*). These include three occurrences of ONE'S SELF, as against five occurrences of ONESELF, listed separately.

The 67 occurrences of indefinite (human) ONE'S/ONESELF are spread over 38 of the 500 sample/source passages which constitute the Corpus, as compared with 65 occurrences spread over 47 samples in the LOB Corpus. The occurrences of ONE'S/ONESELF lend themselves to a twofold classification, distinguishing between anaphoric ONE'S (with co-occurring, antecedent ONE) and "absolute" ONE'S. In glosses of both monolingual dictionaries and pedagogical grammars attention is focussed on the use of anaphoric ONE'S, but in the Corpus it is less common than absolute ONE'S (24 occ., 36%, as against 43 occ., 64%). Anaphoric ONE'S is, in each case, coreferential with its antecedent ONE, as exemplified in the illustrative sentences of the dictionaries and grammars. However, *his* and *himself* as correlatives of *one*, to which they draw notice, as alternatives to ONE'S/ONESELF in AmE, are rare in the Corpus texts (3 occurrences as against 24 of anaphoric ONE'S).

The realizations of absolute ONE'S may, it is suggested, be further divided into two main groups according to the syntactic environment in which ONE'S occurs, roughly coinciding with a non-finite and a finite clause structure, coreference with the latent notional subject (indefinite and human) of the non-finite verb generally

characterizing the former, and an identity of ONE'S independent of the other sentence constituents, the generic pronoun, being its interpretation in the latter.

Of the 43 occurrences, 28 occur in a non-finite clause or nominalization with verbal force, where a notional subject (human with generalizing function) is implied, with which, with possibly one exception ([3] below), ONE'S is coreferential. Of the other 15 occurrences, 13 are instances of ONE'S, the generic pronoun, variously interpretable as having general reference (= my, your or anyone's), as a generalization/objectification of the writer's own experience or opinion, or as a (contextually determined) substitute for the first person pronoun.

The remaining two realizations of ONE'S in a finite clause do not occur in a syntactic environment in which coreference with a latent notional subject can be inferred. On the other hand, situational/semantic constraints inflect the interpretation of ONE'S away from that of the generic pronoun proper (which signals a generalization in which the writer is included) to a quasi-generic ONE'S, i.e. coreference with any or every individual of a group of persons under discussion. In [1] ONE'S does not include the writer, who was the organizer of the course not one of the participants:

> [1] these (discussions) were concerned with the possibility of the death of one parent ... but the possible death of ONE'S own spouse was not overlooked. (J27 1440, D.L.Womble, "Functional Marriage Course for the Already Married")

That this interpretation conforms to AmE but not BrE speech practice is, I would suggest, borne out by the variant definition of ONE (under which ONE'S is subsumed) given in W-3 (1961): "an individual of a vaguely indicated group", a semantic variety of ONE which does not appear in BrE dictionaries.

Other points of divergence between AmE and BrE exemplified in the Corpus texts were a) the simultaneous coreference of ONE'S with an explicit antecedent and a latent notional subject as in:

> [2] everybody has followed the New England pattern of segregating ONE'S children into private schools (G17 0950, R.Stewart, "A Little History, a Little Honesty")

and b) the possible non-coreference of ONE'S in a non-finite clause, with the latent notional subject of the infinitive as in:

> [3] To remove an insect from ONE'S ear warm water should be inserted. ... Another way to remove small objects from the eye was to have the person look cross-eyed (F26 1010, Amy Lathrop, "Pioneer Remedies from Western Kansas")

Evidence from American and British dictionaries seem to corroborate this divergence in usage. Cf. the glosses: AmE – "pull ONE'S weight" and "pull ONE'S leg"; BrE – "pull ONE'S weight", but "pull a person's or someone's leg".

[1] & [2] were among the instances of ONE'S in the Brown Corpus rated as conforming to the norm of American speech practice by an American linguist

whom I consulted. The findings of a small scale elicitation enquiry with regard to the use of ONE'S by AmE and BrE speakers, lend further support to the hypothesis that AmE permits relaxation of certain constraints which are a feature of BrE, as illustrated in [1] – [3].

References

Devons, N. 1985. FREQSUCON. *ICAME News* 9: 20-22.

Devons, N. 1986. Observations on lexicographic treatment of ONE and the approach adopted in FREQSUCON. *ICAME News* 10: 31-32.

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

LOB Corpus. 1978. Lancaster-Oslo/Bergen Corpus of British English for use with digital computers.

W-3. 1961. *Webster's Third New International Dictionary of the English Language.* Springfield, M.A: Merriam.

Acknowledgement

I am very much indebted to Crawford Feagin of Washington, D.C., the American linguist whom I consulted.

Exploring the LDB: Noun phrase complexity and language variation

Pieter de Haan
University of Nijmegen

Since Aarts (1971) no attempt has been made to demonstrate the relationship between the distribution of noun phrase type and text variety, although in Aarts it was hinted that such relationship exists. It has recently become possible to carry out research of this nature relatively easily, because of the existence of the *Linguistic Database* (LDB) and the loglinear analysis, which makes a detailed and complex analysis of corpus material possible.

Based on the investigation of the written part of the Nijmegen corpus (approx 120,000 words), currently available in the LDB, it was shown that there exist complex relationships between noun phrase type, noun phrase function and text variety. The noun phrase types distinguished ranged from basic to complex. Three noun phrase functions were examined: *subject, direct object* and *prepositional complement*. The Nijmegen corpus contains equal numbers of fiction and non-fiction samples. A loglinear analysis of the material showed that the distribution found is highly significant.

One of the great advantages of the LDB is that the exploration of the data can be carried out interactively, which makes it possible to examine actual corpus

examples during the procedure. This may result in the formulation of new hypotheses, which can then be tested by investigating further material. It is thus possible to refine previous hypotheses. Apart from this, the LDB can be used to carry out frequency counts and to produce frequency tables of specific structures.

Both uses of the LDB were involved in the project discussed. First a frequency table was generated. The figures in this table were used as the input for the loglinear analysis. The results of the analysis gave rise to a further inspection of the LDB, during which some specific details of differences between the distribution of certain noun phrase types in the two text varieties were examined in the LDB.

Reference

Aarts, F.G.A.M. (1971). "On the distribution of noun-phrase types in English clause-structure", *Lingua*, 26:281-293.

Interaction in syntactic corpus analysis

Theo van den Heuvel
University of Nijmegen

Future activities within corpus analysis can be expected to lean heavily on a syntactic analysis of some sort: a projection of utterances upon (sets of) syntactic interpretations. A fully automatic syntactic corpus analysis is impossible to the degree that it makes use of "general knowledge of the world". On the other hand, there are strong arguments for automation such as the sheer complexity and extent of the work involved and the need for consistency. Therefore syntactic corpus analysis will normally proceed in interaction between linguist and computer.

This paper presents a framework for a methodological discussion of man-machine interaction in syntactic corpus analysis, abstracting from the actual installation, formalism and grammatical theory used. It explores the possibilities and limitations of automatic and computer-aided syntactic analysis. It attempts to arrive at a scenario of syntactic corpus analysis with minimal effort from the side of the linguist.

The Helsinki Corpus of English Texts: Diachronic and Dialectal – Report on work in progress

Ossi Ihalainen, Merja Kytö, and Matti Rissanen
University of Helsinki

1. The diachronic corpus

The purpose of our diachronic corpus is to provide a tool for empirical research and for analysing variation at the past stages of the English language. It will consist of a collection of texts and text extracts in a machine-readable form dating from the eighth to the early 18th century. The corpus is intended mainly for syntactic and lexical studies, but we hope that students of phonology, morphology and even style may find it useful.

Our corpus is divided into two parts: "the basic corpus" and "the supplementary corpus". The basic corpus is collected and classified systematically, and that will be the version routinely referred to, distributed and worked upon. The suplementary corpus consists of computerized text material collected by individual scholars for their own research work and it is based on their special needs and interests. The supplementary material should feed new items into the basic corpus, and thus make it more varied and better serviceable in the future.

The size of our basic corpus will be approximately 1.5 million words. The length of the extracts varies, but we aim at a minimum length of c. 5,000 words and a maximum of c. 10,000 words from each text. Because the basic aim of the corpus is to serve the purpose of syntactic study, it mainly comprises prose extracts, although a selection of Old and early Middle English poetry is included as well as samples from late ME and early ModE verse drama.

It is our aim to make our corpus as balanced and representative as possible with respect to chronology, dialect, text type and style. Of the other parameters, the most important are the relationship of the text to spoken language, the relationship of the text to a foreign original, and the description of the author and the audience. At present, there are nineteen reference categories in our coding scheme.

We hope to have the writing of texts finished by the end of 1986, the coding completed by the following spring, and the first version of the entire corpus ready for distribution to scholars in 1988. Although we are unable to send computer data abroad at the moment, scholars are welcome to visit our project and consult our data in Helsinki.

2. The dialectal corpus

Our aim is to compile a half a million word corpus of dialectal British English. The texts are tape-recorded interviews with elderly speakers of traditional rural vernacular from the 1970's. The interviews are transcribed orthographically. At the

present moment about 130,000 words of the Somerset, Devon, Suffolk, Cambridgeshire and Yorkshire dialects have been stored. A system of tags that would be suitable for the study of dialectal syntax is being developed. The extent to which the text can be tagged automatically will be seen in the near future.

To accompany the corpus a number of grammatically analysed sample texts stored in dBASE III files will be created to illustrate the special character of each dialect. The user will obtain a general picture of the dialects in the corpus by asking the database system such basic questions as "What are the main dialectal features?", "What does the past tense form of *to be* look like?", "What sort of verb-initial syntax does this particular dialect display?", etc. There will be memo files attached to the actual records which will comment in greater detail on the language, and provide the user with further references (i.e. places in the corpus and in the literature).

3. Compiling the corpus: Practical aspects

The texts of the diachronic corpus have been selected and the parameters discussed by the scholars and research students working on this material; the dialect informants have been interviewed and the tapes of the interviews transcribed by those concentrating on the dialectal part. The team responsible for keying in the texts consists of three or four advanced students of English Philology, employed on part time basis.

In order to ensure easier word-processing and safer work files, the keying in is done onto micro-computer diskettes (the machines so far used are the Kaypro 2x and 4, and the Olivetti M24). From the diskettes the material can be transferred to the mainframe (Burroughs 7800, currently). The aim is to make the corpus available for use on both micro-computers and the mainframe. To make the study of the individual texts or text groups as flexible as possible, the one-file version of the corpus will be accompanied by a version in which each text remains in a work file of its own.

The basic principle in keying in the material is to preserve as much as possible of the original text within the character set available. In a corpus covering the different stages of English, the editorial principles and typographical conventions necessarily vary from one text to another. A one-file machine-readable corpus, however, calls forth standardization within the limits set by the search programs currently available. Among the features taken into consideration are runes, foreign language, font other than the main font, emendations, editorial comments, headings, superscript and accents derived from foreign spellings. The Old and Middle English special characters will be replaced by a combination of an asterisk followed by the nearest Modern English character (u.c. *ash* = *A; l.c. *ash* = *a etc.). The set of conventions will also be applied to texts received in magnetic form from the Oxford Text Archive or other sources.

The mainframe programs used so far to process the diachronic material are the

LINUS devised for the Burroughs machine by Dr. Kimmo Koskenniemi at the University of Helsinki and the OCP supplied by the Oxford Text Archive. The COCOA format of the OCP will be adopted for encoding the parameters.

So far som 1,321,000 words have been keyed in from the diachronic texts (the 465,000 words taken from the Toronto Old English Corpus included); the number of the words keyed in from the dialect material is 130,000.

The tagged LOB Corpus: Description and analyses

Stig Johansson
Knut Hofland

The LOB Corpus is now available in two tagged versions. Also available is a homograph-separated concordance, sorted by key word and tag (and including frequency information). See later in this issue of the *ICAME Journal*

Examples were given from the analysis of the tagged LOB Corpus on the following levels:

- tag frequencies
- word frequencies
- tag combinations
- word combinations

The frequency of the main word classes is broadly comparable with the Brown Corpus, but there are considerable differences between the two main category divisions of the corpora (informative vs imaginative prose). New word-frequency lists have been produced, sorted alphabetically and by rank. The new lists are homograph-separated (but not lemmatised).

Tag combinations have been studied in two main ways. New tag-pair statistics have been calculated and have been fed into the revised version of the word-tagging programs developed at Lancaster. A second type of analysis looks at the frequency of tag combinations at the beginning and the end of sentences. List of such sequences were used to campare the distribution of definite and indefinite noun phrases. The distribution in general agrees with what would be expected from the well-known tendency to present information in the order given-to-new. But noun phrases which occur as complements of prepositions deviate from the main pattern. There is also an interaction with length. Longer phrases (only premodifiers were counted) are relatively more frequent at the end. Moreover, the relative frequency of the indefinite article increases the more premodifiers there are.

Lists of recurrent word combinations are derived from the tagged corpus, similar to those made for the Brown Corpus by Kjellmer (see below). A preliminary scheme for the identification of relevant word combinations was presented.

The detailed results of the analysis of the tagged LOB Corpus will be published in a forthcoming book.

Aspects of English collocations

Göran Kjellmer
University of Gothenburg

Fairly long collocations (≥5 words), viz. collocations of a somewhat fossilised nature, are particularly at home in the more formal genres of the Brown Corpus, those sometimes referred to as "informative". Also, collocations in general are more frequent in formal/informative genres of text than in informal/imaginative, probably because writers of the former type of text are more likely to fall back on stereotypes, ready-made patterns, than are writers of the latter type of text, where originality is more of a virtue. And finally collocations in ALL kinds of text are essential, indispensable elements, elements that are often neglected as the material with which our utterances are made.

Lexical resolution

Antoinette Renouf
University of Birmingham

This paper examined the characteristics of word forms as they are revealed in text corpora of differing sizes, with particular attention to the ways in which a larger corpus supports or modifies the information supplied by a similar, but smaller, one. The data sources which were consulted both form part of the Birmingham Collection of English Text. The smaller one consists of about 7.3 million running words, and the larger one of about 13 million running words, and for convenience they are referred to in the paper as the "Main Corpus" and "Reserve Corpus" respectively. In its auxiliary role, the Reserve Corpus has a number of effects. Firstly, it confronts the researcher with instances of rarer lexical items which do not occur at all in smaller amounts of textual data, and which traditionally do not receive the lexicological attention which they may merit. Among the word forms in this category are, for example, those which refer to the natural world, and the less frequent forms of a lemma.

Where word forms already occur in the smaller corpus, the Reserve Corpus brings further evidence of their meanings and patterns of behaviour. Sometimes this supports the initial impression; in some cases, it introduces counter examples, which must then be taken into account. Occasionally, it highlights oddities of usage, such as ambiguity, which might otherwise go unnoticed. A large corpus also allows the researcher to look beyond the word form to the lemma. In the Reserve Corpus, for example, lemmas can seem to collocate with each other in various ways.

No definite conclusions can be drawn about the size of the corpus which would be ideal for lexicological purposes. However, it seems that the number of new word

forms occurring decreases as a corpus grows, so that more evidence will be available for more of the word forms in a larger corpus than in a smaller one.

Evidence from the LOB Corpus against the "grammatical"/"ungrammatical" distinction

Geoffrey Sampson
University of Leeds

The majority of computational linguists develop systems which analyse NLs using some type of generative grammar which defines a clearclut class of well-formed sentences. But computational linguists who work with corpora of authentic NL material often doubt the validity of any clearcut distinction between grammatical and ungrammatical sequences. Statistics on the distribution of different types of noun phrase in a 40,000-word sample of written English are used to show (i) that there is a continuous gradient from very common to very rare constructions, and (ii) that alternative constructions grow more numerous at lower frequency-levels in a regular fashion which implies that a significant proportion of grammatical constituents in a text will belong to a extremely rare types.

Carry-on signals in English conversation

Anna-Brita Stenström
Lund University, Sweden

This paper concentrates on the use of *right, right o, that's right, all right, that's all right* and *it's all right* as interactive devices in conversation. The data was collected at the Survey of English Usage, University College London and covers 89 spoken texts of 5,000 words each. Following the identification of discourse function, all the items, with an indication of the most characteristic features, were stored as a database for further processing with dBaseII.

The overall distribution of the 490 instances in the corpus was as follows:

right	272	55%
all right	117	24%
that's right	79	16%
that's all right	9	2%
right o	8	2%
it's all right	5	1%
TOTAL	490	100%

The starting-point for identifying their interactive role was the position in a turn. An item could either constitute a turn of its own or occur at the beginning, within, or at the end of a turn:

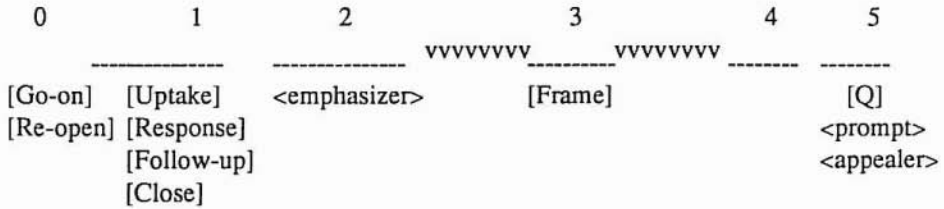
Each item was found to occur in more than one position, usually with a different function in a different position.

Contrary to Sinclair & Coulthard (1975) and Edmondson (1981), I included the level of turn in the model of analysis. The way the model captures some of the uses of *right* and *all right* as carry-on signals is illustrated below:

			TURNS	Moves	acts
A: RIGHT -			{ }	[ ]	< >
Nic turned up	T	EXCHANGE 1		[ ]	< >
B: Nic Kay	R		{ }	[ ]	< >
A: yes	A		{ }	[ ]	< >
B: RIGHT	N		{ }	[ ]	< >
	S				
A: did you talk to him	A	EXCHANGE 2	{ }	[ ]	< >
A: No	C		{ }	[ ]	< >
I was away	T				< >
B: ALRIGHT	I		{ }	[ ]	< >
	O				
but you could have phoned	N	EXCHANGE 3		[ ]	< >
RIGHT					< >
A: of course			{ }	[ ]	< >

The first *right* is a [Frame], i.e. a move that marks a boundary in the discourse and signals the transition between two stages. The second *right* is a [Follow-up], a move that evaluates the previous response and terminates the exchange. *Alright* is an [Uptake] by which the speaker links his next move in the same turn with the previous speaker's move. And the third *right* is a <prompt>, an act which transforms the statement to which it is attached into a request for confirmation.

This is how position in the turn was typically related to discourse function:



Right, the most common carry-on signal, served all the functions but was most frequently used as a [Follow-up] and a [Response]. *All right* was characteristically used as a [Frame]. *That's right* was the typical <emphasizer>, e.g. /Y^ES# /that's R^IGHT#, i.e. a secondary act emphasizing a preceding primary act. *It's all right* and *that's all right* served as [Responses] to <apologies>. *Right o* finally, was either a [Response], a [Follow-up] or a [Go-on]. The [Close], by which speakers end a conversation, was realized only by *right* and *all right* and occurred almost without exceptions in telephone calls.

Considering that 11 of the 89 spoken texts consisted of telephone conversations and that 57% of the carry-on signals occurred in telephone calls, it can safely be stated that the use of at least *right, all right* and *right o* in British English (not only functioning as [Closers]) is highly related to the speakers' mode of communication.

Taking a new look at word class tags

Jan Svartvik
Lund University

The project Text Segmentation for Speech (TESS) has as its immediate aim to study the segmentation of English speech into tone units and, on the basis of the insight derived from the study of the prosodic and grammatical properties of such units, to set up some of the rules which govern the natural segmentation of spoken discourse. The long-term aim is to make a contribution to the understanding of the principles of human "chunking", which can be applied to, for example, the improvement of the quality of text-to-speech conversion. It is a basic tenet in our approach that the division of connected speech into tone units is an important element in human speech processing.

An important element in our approach is the automatic grammatical analysis of tone units by means of a parser. Automatic analysis of genuine texts requires a close interplay between the different levels of grammatical analysis, and it is therefore necessary to revise, from time to time, the categories on one level in order to achieve a better result on another level. During the past year a new set of word-class tags has been introduced. Yet I would prefer not to consider even the current set of word-class categories as finalized but rather open to revision throughout the research process.

The paper outlines the principles behind the new tagging system and gives a list of current tags, which are more "delicate" and hence more numerous than the old set of tags.

Program distribution and networking within ICAME

Knut Hofland
Norwegian Computing Centre for the Humanities
Bergen

Program distribution

At the recent ICAME conference in Amsterdam there was a special session on the exchange of programs for use in corpus linguistics and related fields. As a result of this discussion, the Norwegian Computing Centre for the Humanities has set up a service for the collection and redistribution of programs. There is no restriction on type of programs, programming language or operating system. The programs will be distributed on floppy disk, tape or via network (see below). Users that have programs they are willing to share with others are requested to contact the Centre in Bergen. A list of available programs can be obtained via net or from the Centre in Bergen.

ICAME network mailing list

An electronic mailing list of ICAME users has been set up. This list will be used to distribute information from the Centre in Bergen between the issues of the ICAME Journal. If you want to be added to this list, or have information that you want to distribute to the people on this list, please contact the coordinator at the address given below.

ICAME network server

To facilitate the distribution of information and programs, a network server has been set up at the EARN/BITNET node in Bergen. This server can be reached from any network that has a gateway to EARN/BITNET like Uninett, Janet, Arpa, Csnet etc. The server contains information about the material available, some text samples, an ICAME bibliography, programs and documentation, and network addresses. The server can be contacted in two ways:

a) via interactive messages (only EARN/BITNET)

Example from the IBM VM/CMS environment:

TELL FAFSRV AT NOBERGEN help

will .give the following answer

```
*> NAVF/ICAME Bergen 1 Apr 1987 09:14:11
*> Available commands:
*> HELP - Send this information
```

*> SEND - Send list of available files
*> SEND fn ft - Send specified file
*> QUERY msg - Store msg to server operator
*> You may also send mail to server operator
*> End of NAVF/ICAME Server Help Info Bergen

TELL FAFSRV AT NOBERGEN send icame netaddr

will send you the list of names on the ICAME electronic mailing list

*> NAVF/ICAME Bergen 1 Apr 1987 09:14:53
*> The file ICAME NETADDR has been sent to you
PUN FILE 1574 FROM FAFSRV COPY 001 NOHOLD

b) via mail

The commands to the server are given in the subject line. Only one command is available in each letter at the moment.

Example:

Date: 12 Mar 87 16:45 -0100
From: Stig Johansson <h_johansson%use.uio.uninett@cernvax>
To: fafsrv@nobergen
Subject: send test boo

How to transfer MS-DOS programs

The MS-DOS programs (.COM or .EXE files) are stored as 8-bit bytes. These files can be transferred between some EARN/BITNET sites, but not all and not via gateways to other networks. To test this, request the file TEST COM. Transfer this file from your local host with the Kermit program. Make sure to set the file type to binary with the command SET FILE BINARY to the host Kermit. Run the program on your PC. You should then see all the ASCII characters displayed.

Another way to transfer binary files is to encode the file as a file of 7-bit printable characters, transfer the file, decode the file back to an 8-bit file. Files encoded in this way have the extention .BOO. To test this transfer, request the files FROMBOO PAS and TEST BOO. FROMBOO PAS is a Turbo Pascal program that decodes a .BOO file. Transfer these files to your local PC, this time as text files (default to the host Kermit). Strip off the mail headers if you have requested the files via mail. Compile the FROMBOO.PAS program and run the program. Give the name of the input file TEST. The file TEST.COM will now be generated. Run the program and you should then see all the ASCII characters displayed.

The program that decodes a file is named TOBOO.EXE or TOBOO.BOO. This can be used if you want to transfer programs via net to Bergen.

In the future other decoding and compression techniques may be used.

Server

EARN/BITNET: FAI'SRV@NOBERGEN
JANET: FAI'SRV%NOBERGEN@UK.AC.RL.EARN
ARPA: FAI'SRV%NOBERGEN.BITNET@WISCVM.WISC.EDU

Coordinator

EARN/BITNET: FAI'KH@NOBERGEN
JANET: FAI'KH%NOBERGEN@UK.AC.RL.EARN
ARPA: FAI'KH%NOBERGEN.BITNET@WISCVM.WISC.EDU

New material

The Augustan Prose Sample was presented in *ICAME News* 4 (1980). It is introduced again below, as there is now available a diskette version. The CHILDES project will be presented more fully in a later issue of this journal.

The CHILDES Project

The CHILDES project has set up a system for the exchange of child language data: The *Chi*ld *L*anguage *D*ata *E*xchange *S*ystem. The work is directed by Brian MacWhinney, Carnegie-Mellon University, who also edits the CHILDES newsletter. Altogether the CHILDES data base includes 21 corpora of parent-child interactions from English-speaking children. Some of these corpora have only one subject; others have much larger numbers. There are also corpora from several other languages. About one half of the data has been reformatted into a standard transcript format specified by the CHILDES project. There are plans to circulate data (120 megabytes) on CD-ROM. The funding has come from the National Science Foundation and the John D. and Catherine T. MacArthur Foundation. For more information, write to: Brian MacWhinney, Department of Psychology, Carnegie-Mellon University, Pittsburgh, Pennsylvania 15213, USA.

The Augustan Prose Sample

The Augustan Prose Sample is a machine-readable data base made up of prose selections amounting to 80,000 words drawn from texts published in England during the period 1675-1725.

The Sample consists of a sequence of 1650 sentences drawn from 52 selections by such authors as Arbuthnot, Burnet, Evelyn, Locke, Mandeville, and Temple, among the better known, and Burthogge, Dunton, Hoadly, Prideaux, and Winstanley, among the more obscure. The selections range from 352 to 4298 words in length and average 1522 words. For positive identification and to permit sentences to be used individually or in random combinations, each sentence is preceded by a block containing an identification number, a selection code, the date of original publication of the selection, the sequence number of the sentence in the selection, and the number of words in the sentence, thus:

0812-ET/1699:006(046) HE WAS NEVER VERY HEALTHY, NOR TOO SICKLY;...

To achieve representativeness, it was decided to choose texts from authors representing a cross-section of the publishing activity of the period – what was likely to be read by the average literate gentleman in the coffee houses of London – rather than the outstanding writings of literary writers. In particular it seemed

desirable to include a wide variety of forms and topics, as well as writers at various stages of their careers. The original plan called for one selection per year, but for a variety of reasons this proved impossible to achieve. Hence some years are over-represented and others are blank. For example, 1702 has five selections (from writers aged between 30 and 76), whereas 1700, 1701, 1703 and 1704 have none. Nonetheless, it is probable that the Sample adequately reflects the variety of works published at the time.

To produce an accurate text requires access to the earliest printings or facsimiles or scholarly editions. Such were not always available but it is doubtful that the integrity of the Sample has suffered as a result. Nor is the Sample a verbatim reproduction of the texts from which it is constructed. For ease of programming, spellings were regularized to the American standard (although an original-spelling version is available). Similarly, punctuation has been slightly simplified – periods are used only for end of sentence pointing – and the whole is in upper-case letters with dollar signs to indicate proper names.

The Sample has been available in a tape version for mainframes. It will now be available in a diskette version for personal computers. Full documentation, including the entire text, selected statistics and instructions, may be secured on request at cost. To receive an order form for tape, disks or book write to: Louis T. Milic, Department of English, Cleveland State University, Cleveland, OH 44115.

The Alice Concordance

The Alice Concordance is a condensed concordance to the books *Alice's Adventures in Wonderland* and *Through the Looking-Glass* by Lewis Carroll.

It is complete except for the omission of words occurring more than one hundred times, of which there are ninety-seven. This results in a manageable and useful volume containing just over twenty-three thousand key words with contexts.

The method used for constructing the contexts is designed to produce a useful context in only a small space. It does this by taking punctuation into account, and overall is remarkably successful. Here are some sample entries:

Snowdrop 1L647 you pulled Snowdrop away by the tail just as I
Snowdrop 12L285 "Snowdrop, my Pet!" she went on, looking over

These tell us that the word "snowdrop" occurs in the contexts shown, in chapters 1 and 12 of *Looking-Glass*, as words 647 and 285 in those chapters respectively. An introduction fully describes the content and layout of the entries.

As a bonus, *The Alice Concordance* also contains two vocabulary listings. One of them lists the words alphabetically and the other lists them in descending order of frequency.

For copies write to: The Alice Concordance, The Language Laboratory, The University of Adelaide, GPO Box 498, Adelaide SA 5001, Australia. The price is $A9.50, plus postage.

Material available through ICAME

The following material is currently available on computer tape from Bergen through the International Computer Archive og Modern English (ICAME):

Brown Corpus, text format I (without grammatical tagging): 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, text format II (without grammatical tagging): This version is identical to text format I, but typographical information is reduced and the line division is new.

Brown Corpus, KWIC concordance (also on 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.

LOB Corpus, untagged version, text: 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 (also on 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: 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: 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 (also on 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.

London-Lund Corpus, text: 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: 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: A complete concordance for the remaining 53 texts of the London-Lund Corpus (text categories 4-12).

Melbourne-Surrey Corpus: 100,000 words of Australian newspaper texts (see the article by Ahmad and Corbett in this issue of the journal).

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 from Bergen.*

Some of the material is being prepared for distribution on diskette (see the order forms). There are also plans to distribute material on CD-ROM.

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, however, recommended to consult J. Svartvik & R. Quirk, *A Corpus of English Conversation* (see above).

Information about ICAME and order forms can now also be obtained from:

Oxford Text Archive, Oxford University Computing Service, 13 Banbury Rd., Oxford OX2 6NN, England

Humanities Research Center, Brigham Young University, 3060 JKHB, Provo, Utah 84602, USA

These centres also assist in distributing material. All order forms are sent to Bergen.

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 English, 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