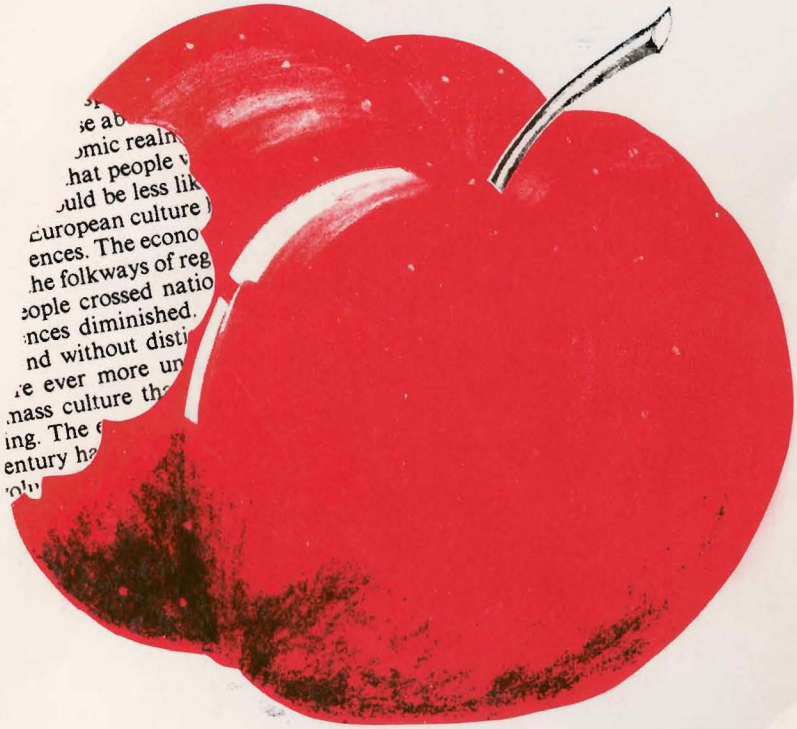
ICAME Journal

Computers in English Linguistics

No. 17

April 1993



International Computer Archive of Modern English

Norwegian Computing Centre for the Humanities

ICAME Journal

Computers in English Linguistics

No. 17

April 1993

International Computer Archive of Modern English

Norwegian Computing Centre for the Humanities

International Computer Archive of Modern English (ICAME)

ICAME is an international organization of linguists and information scientists working with English machine-readable texts. The aim of the organization is to collect and distribute information on English language material available for computer processing and on linguistic research completed or in progress on the material, to compile an archive of English text corpora in machine-readable form, and to make material available to research institutions.

The Norwegian Computing Centre for the Humanities in Bergen, Norway, acts as a distribution centre for computerized English-language corpora and corpus-related software. It publishes the *ICAME Journal* (previously *ICAME News*) and maintains an electronic information service (for details, see p. 139 in this issue). Conferences have been arranged since 1979.

ICAME ADVISORY BOARD

For information on ICAME, contact: Norwegian Computing Centre for the Humanities, Harald Hårfagresgt. 31, N-5007 Bergen, Norway.

Editors of the *ICAME Journal*:
Stig Johansson (Oslo)
Anna-Brita Stenström (Bergen)

Contents

ICAME services:

The marking of cohesive relationships: Tools for the construction of a large bank of anaphoric data

Roger Garside
University of Lancaster

Abstract: In the creation of large text corpora, the quick and accurate manual annotation of text is often important. This paper describes the design of a task-oriented editor (XANADU) with which a team of analysts can mark the main cohesive links in a text, such as between a pronoun and its antecedent or between an ellipsis site and the ellipted material. The editor has been designed to allow convenient use of window and mouse technology to simplify and speed the analyst's task. XANADU has been used to annotate some five thousand sentences during its development, and this usage has prompted a number of modifications to the original design.

1. Introduction

Nowadays the importance of corpora in Natural Language Processing is becoming increasingly acknowledged, both for the construction of natural language processing systems and for their evaluation. For many requirements the corpora need to be annotated with syntactic or semantic information, and this is often a difficult and time-consuming process if it is to be accurate and cover a substantial amount of data. The annotation process needs to be an appropriate division of labour between machine and manual processing. In our research at the University of Lancaster we have been concerned with the optimal interaction between manual skills and automatic processing, and have developed a series of 'intelligent editors' to aid in the annotation of texts. Early work was focussed on the syntactic annotation of texts, and this is briefly reviewed in the next paragraph. More recently, we have been working on the construction

of a corpus annotated to show the main cohesive or anaphoric relationships in the texts, and for this we have developed a task-oriented editor called XANADU, which forms the main subject of this paper. This editor exploits window and mouse technology to allow the user quickly and efficiently to mark cohesive relationships.

The building of annotated corpora has been going on at Lancaster since 1980, in a group called UCREL (Unit for Computer Research on the English Language) comprising members of the Departments of Computing and of Linguistics and Modern English Language. Early work included the part-of-speech annotation of the million-word LOB (Lancaster-Oslo/Bergen) Corpus, and the construction of various 'tree-banks' of a few thousand or tens of thousands of sentences, annotated with detailed constituent labels to show the surface form of the parse tree for the sentence. Since 1987 work has been directed to building a further corpus with the principal syntactic constituents labelled. This work, funded by the IBM Thomas J. Watson Research Center, Yorktown Heights, and the IBM Scientific Centre, Winchester, has as its main aim the training and subsequent evaluation of a robust probabilistic grammar of English for use in speech recognition by computer. Since this main aim necessitates the construction of a very much larger annotated corpus than before (at least several million words), it is necessary to opt for a simpler form of marking, called 'skeleton parsing', which can be applied quickly and consistently by a larger number of analysts. For this a special-purpose editor was developed which enables the analysts to input the annotation as they parse the sentence; a peak rate of up to a sentence (on average twenty words long) per minute has been achieved. The development of syntactic annotation systems at Lancaster, and the technique of skeleton parsing, are described in detail in Leech and Garside (1991).

The resolution of pronoun references is important in natural language understanding, machine translation, automatic abstraction of texts, etc; we therefore felt that it would be useful to collect a corpus of texts marked up to show the targets of pronoun references. In late 1989 it was agreed between the UCREL and Yorktown Heights teams that we begin construction of a corpus marked to show explicitly a variety of anaphoric or, more generally, cohesive relationships found in the texts. With funding from IBM this has gone on in parallel with the continuing skeleton parsing, with on average some three-quarters of the analysts' effort going into the latter. The concept of what to attempt to mark, the annotation scheme to be used, and the editor used to capture and check the annotation have been developed in parallel, allowing feedback

on difficulties encountered by the analysts in using any particular version of the system. In the last six months the development work has been essentially completed, and to date (Summer 1992) more than 5000 sentences have been annotated in accordance with the latest recension of the notation system.[1]

This paper describes the results of this development process, with some discussion of the choices made and the reasons behind them. Section 2 gives an overview of the notation used to mark anaphoric and other cohesive features in a text. Section 3 describes the basic procedures employed in marking the texts, and the XANADU editor. Section 4 describes in detail some of the special features of the XANADU editor for this type of annotation, and the changes made to the features in the light of the analysts' experience with the difficulties of performing the required tasks. Section 5 discusses some of the subsidiary, but important, aspects of the analysis procedure, such as mechanisms for quality control. Section 6 reports on results to date and possible future developments. Work is continuing to expand the size of the annotated corpus, and we are beginning to consider the modification of the XANADU editor to allow its use in other text annotation areas, since we believe that it is important to continue the development of efficient editing tools if progress is to be made in corpus annotation.

2. *The notation*

The plans for generating a corpus of texts with explicit marking of cohesion assume a similar goal to that of the corpus with explicit syntactic marking; a large quantity of marked text has to be produced (several million words for the syntactic corpus, at least several hundred thousand words for the anaphoric corpus), so speed of human analysis is important, but not at the expense of low accuracy and consistency in marking. The result is a notation scheme which does not attempt to mark all possible theoretical distinctions, and in fact tends to be theoretically fairly neutral, although it is influenced by the scheme described in Halliday and Hasan (1976). An over-riding principle to be borne in mind by the analysts using the scheme is that a feature should be marked only if they are fairly sure of it.

The current marking scheme started from a first draft in 1989 of the types of feature to be marked and the annotation to be used. This was elaborated and guidelines added, and was then tested by application to corpus texts by the UCREL team of analysts. This development cycle has been iterated several times, in attempting to resolve a tension

between what it would be theoretically interesting to mark and what it is possible to capture with consistency, given the required volumes of data and speed of marking. Although the notation is inserted automatically by the software described in the next section, it has been designed so that it is reasonably easy to read, with the commoner structures represented by simpler markings; it could of course be translated into a standard notation, such as that currently being developed by the ACH/ACL/ALLC Text Encoding Initiative (Sperberg-McQueen and Burnard 1990).

The remainder of this section gives a brief overview of the notation. For more details of the notation, the linguistic principles behind its construction, and discussion of the guidelines used by the analysts in marking up a text see Fligelstone (1991, 1992).

The basis of the notation is that, in a typical anaphoric link between a proform and an antecedent, the antecedent is enclosed in brackets and given an index number which is unique within the text, and the proform is preceded by a symbol indicating an anaphoric referential link to that numbered antecedent, thus:

(6 the married couple 6) said that <REF=6 they were happy with <REF=6 their lot.

Here the character '<' indicates to the human reader a preceding antecedent (ie the link is anaphoric rather than cataphoric), although from the computer point of view this would be adequately indicated by the co-indexing of the numbers. A cataphoric link would have the character '>' on the symbol marking the proform. The characters 'REF=' indicate a referential link (as distinguished from substitute forms, ellipsis, etc). Since proforms are nearly always one word long, a length indication (either explicit or implied by brackets round the proform) is unnecessary; in the few cases requiring a proform of more than one word, an explicit length indicator is included, as in:

(7 this week's winner 7) said <REF=7 he had rung (8 <REF=7 his wife 8) and <REF=7,8 they had spoken to <REF=7,8:2 each other.

Here the symbol '<REF=7,8:2' indicates an anaphoric referential link from a proform two words long (*each other*); the link is to a pair of antecedents (*this week's winner* and *his wife*), and this is also indicated in the notation. It is possible to mark, with suitably placed question marks, doubt about the extent of an antecedent and uncertainty about a proform linking to a particular antecedent. A recent addition to the notation allows a distinction between multiple reference and alternative reference; while the notation '<REF=7,8:2' means a reference to ante-

cedents numbered 7 and 8, the notation '<REF=7/8:2' would mean a reference to either antecedent 7 or antecedent 8. More complicated examples are possible (though rare in the corpus to date); thus '>REF=1,5/6,?22' would mean a cataphoric reference to 'antecedent' 1, either 5 or 6, and probably 22.

Other types of cohesive markings which are indicated by variations on this notation include:

a. Substitute forms, including *ones*, *others*, pro-verbs such as *do so*, etc:

Asked if I were (9 going 9) , I said I would <SUBS=9:2 do so.

b. Ellipsis, such as in gapping, but only if the ellipted material can be recovered from the text (perhaps with some morphological adjustment):

Asked if I were (9 going 9) , I said I would <ELIP=9 .

Notice here that a point in the text, rather than a group of words, is indicated by the symbol '<ELIP=9'. This has important consequences for the design of the software (see Section 4).

c. Meta-textual reference to another point in, or area of, the current text – the current guidelines for this type of marking are extremely limited and tentative, and may be modified in due course:

As has been shown <META above, the experiment was successful.

Here the antecedent of the meta-textual reference is not explicitly marked (but only its direction), so no index number is allocated.

d. Coreferential items which are not proforms are marked by bracketing the items and inserting the same index number, thus:

(10 William Shakespeare 10) was born on St George's Day. (10 England's most famous dramatist 10) was presenting <REF=10 his plays in London by the time <REF=10 he ...

As before, similar modifications to this basic notation allow an indication of a multiple link from one item to several others, or a doubt on the part of the analyst as to the extent of an item or the certainty of a link.

e. Copular relationships between noun phrases, linked by copular verbs or in apposition, have a special notation:

(11 George Bush 11) was {{11 President of the United States 11} .

where the 'extra' brace is used to indicate the direction (in this case leftward) of the copular relationship.

The notation also provides ways of marking features of interest on pronouns (for example whether *you* is singular or plural, whether *we* is 'inclusive' or 'exclusive' of the addressee, etc); with the aim that pronouns would normally be marked for these features irrespective of whether or not they form anaphoric or cataphoric links within the text. Since a text presented for annotation may consist of several independent passages, there is also a symbol to represent a 'cohesion barrier' between one passage and the next, such that no anaphoric link crosses the barrier and a new sequence of indices can commence.

3. *The annotation process*

In order to enhance the usefulness of the marked texts, all the annotation has been done on texts that have already been syntactically marked (in the parallel skeleton parsing project). Details of this marking process are given in Leech and Garside (1991), but briefly such a marked text has a part-of-speech indication on each word, and an indication of the main constituent structure of each sentence. An example (sentence A010/109 from the AP corpus of Associated Press news stories) is:

[N The_AT individuals_NN2 [Tn named_VVN [P as_II [N
targets_NN2 [P of_IO [N the_AT FBI_NNJ probe_NN1
N]P]N]P]Tn]N][V were_VBDR generally_RR keeping_VVG
[N a_AT1 low_JJ profile_NN1 N]V] ._.

Here the part-of-speech tags (attached to the appropriate word by a '_' character) are taken from a set of some 170 'word-tags', and the symbols used to label the constituent brackets are taken from a set of about sixteen. There are additional special annotations for coordination, for grammatical units written orthographically as two or more words (such as *according to*), and for marking 'discontinuous' constituents in certain rare cases. One important aspect of the notation is that analysts are allowed to use unlabelled brackets, in situations where they feel that a syntactic grouping is appropriate but it is not clear which of the available labels (if any) would be the correct one.

Most of the anaphoric marking has been done on the AP corpus, which was divided into units of approximately 100 sentences for the syntactic marking. We have retained this block-size as the basic unit on which the anaphoric marking is carried out. When the AP blocks were selected they always consisted of an integral number of news stories, so there are no anaphoric references across block boundaries. However, a block often consists of a number of news stories, and here

the analyst would make use of the cohesion barrier symbol mentioned above, to separate distinct cohesive passages of text. The process of anaphoric marking of such a text commences when one of the analysts from the team selects a text, and invokes the anaphoric editing program.

The anaphoric editing program is written in C, and runs under X Windows and Unix on several of the Sun work-stations in the Department of Computing at Lancaster University. It currently makes use of the Athena widget set, but we are expecting to rewrite it to run under Motif for the next version of the software. Earlier presentations have described this editor under various names, but it is now called XANADU – this is not (at least at present) an acronym. The program has been through various versions, each of which has been used by the analysts to annotate a number of texts. The principle design aim has been to ensure that the simpler, common types of markings require the minimum number of user actions, if necessary at the expense of more user actions for the more complicated but rarer cases. Another design principle is that most of the annotation should be performed with the mouse, rather than by typing information at the key-board.

An illustration of the situation when an analyst is about to begin the annotation of a text is given in Figure 1. The screen is divided into three main areas:

a. At the top of the display is a portion of the text to be marked, in a window with a scroll-bar to allow the analyst to move through the text. Although the text-file contains part-of-speech markings on each word, and syntactic brackets round the significant constituents, these are not displayed in the window, since they would clutter the text and make it difficult to read it for sense. It is possible to look at the syntactic marking of a particular part of the text if desired (see Section 4). In fact the program was originally designed so that the part-of-speech markings could be displayed along with the words if desired, but the analysts have never felt this to be necessary.

b. At the bottom left-hand corner of the display are three sets of buttons. The main set, at the top, is for inserting the various possible cohesive markings, with one button for each such type of marking – for example anaphoric/cataphoric (proform) reference, substitute form, ellipsis, cohesion barrier, etc. Below these is a set of editing buttons; and below these again is a set of miscellaneous buttons. Clicking with the mouse on one of the cohesion marking buttons brings up a pop-up window containing a set of further buttons, with the appropriate options for this particular marking.

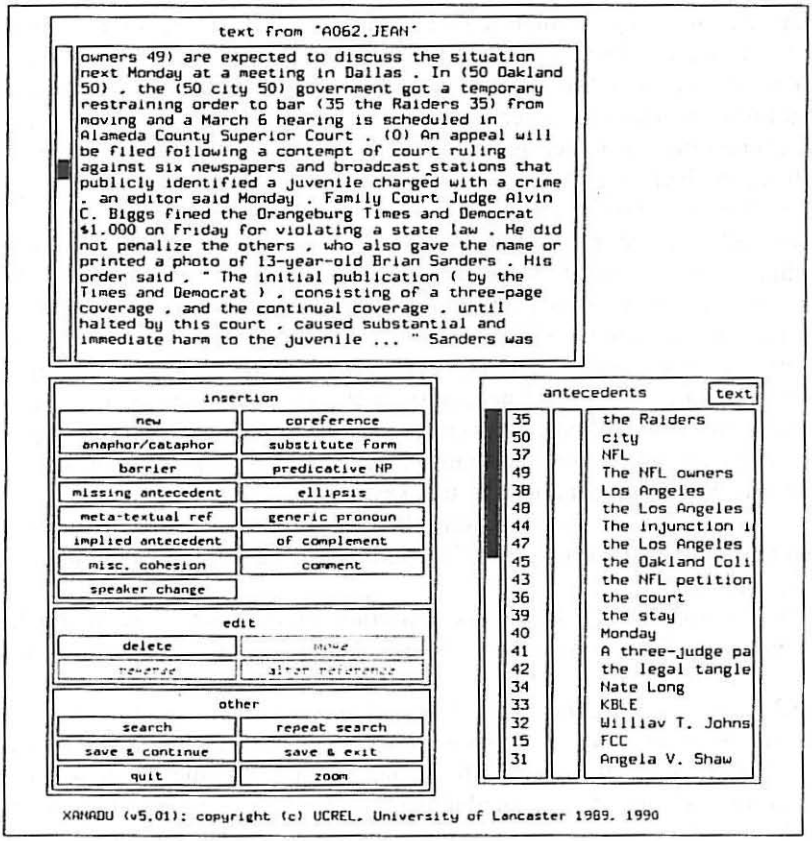


Figure 1

c. At the bottom right-hand corner of the display is a list of all the antecedents so far marked in this text, together with the index numbers allocated to them by the program. In the case of a set of non-proform coreferential items with the same index number, the textually most recent item is the one displayed against the number. The index numbers were, in the early versions of this program, displayed in order of their occurrence through the text, though this is an area where the analysts' patterns of work have led to some modification to the original design (see Section 4).

Consider first the marking of a new antecedent, which should therefore have a unique index number associated with it. The process is for the analyst to click the mouse on the beginning and end of the stretch of text to be marked as the antecedent, and then to click on the 'new' button. This brings up a pop-up menu of options for an antecedent, as shown in Figure 2; this allows such things as uncertainty about the boundaries of the antecedent to be indicated, if desired. Finally the 'confirm' button is clicked – this causes an unused index number to be allocated by the program, and the appropriate marking to be inserted in the text according to the options selected. The marking is displayed at the appropriate place in the text window, and the new antecedent is inserted in the list in the lower right-hand part of the display. It would be possible to eliminate the 'confirm' button, by allowing implicit confirmation by time-out or on detection of the first action for the next marking (ie by clicking on another stretch of text), but this has not been implemented. It was felt that it would be confusing in the early stages of learning to use the program, and it has never been felt by the analysts to be an important improvement since then.

Now consider marking a proform, where we wish to indicate its linkage with a previously-marked antecedent. Here the analyst clicks the mouse on the proform, and then clicks on the 'anaphor/cataphor' button. A single click in the text window indicates that the single word indicated is to be marked. The options menu for the 'anaphor/cataphor' button is displayed, as shown in Figure 3. In a typical simple proform marking, the only further piece of information required is the antecedent number, which is obtained by clicking on the appropriate line in the list of antecedents (after scrolling the antecedents window if necessary; but if the analyst is working through the text in the normal way, the antecedent will usually be currently on display in this window, somewhere near the top). The clicking on the antecedent list is taken as the confirmation for this marking, eliminating an extra button click in the simple case.

There is provision for indicating multiple or alternative references to two or more antecedents (as in the examples in Section 2). Also sometimes the analyst wishes to mark one or more of the reference numbers as uncertain. Both these features are handled with the buttons on the pop-up list. To indicate multiple references the analyst clicks on the 'multiple reference' button, and then clicks on any number of antecedents in the antecedent window, ending by explicitly pressing the 'confirm' button; in order to keep track of what references have been indicated, a small window appears showing the reference numbers

selected so far. Variations on this procedure allow the other rarer types of marking; clicking on the 'uncertain reference' button causes the next reference to be marked as uncertain; clicking on the 'alternative reference' button causes the next pair of references to be marked as alternatives (rather than as cumulative).

Notice that this mechanism requires the antecedents already to be in the antecedent list. In the case of a cataphoric reference, where the proform appears textually before the antecedent, the normal method of working through the text from beginning to end has to be temporarily abandoned, as the antecedent must be marked before the proform can be marked; it is felt that the simplicity of the mechanism and the possibility of error-checking were worth preserving in the relatively rare cases of cataphoric proform reference. As mentioned in Section 2, the marking for a proform includes an indication of direction towards the antecedent. The program makes an automatic attempt to choose the correct direction; but in the case of multiple references, where the reference links may point in both directions, there are buttons to allow an explicit choice by the analyst to override the program.

All the cohesive markings operate in this same general way, with minor variations on what options are available in the pop-up menu for a particular marking. For example, certain types of pronoun marking require pronoun features to be indicated; in this case a button 'pronoun features' brings up a sub-menu of possible features to be marked on this instance of the pronoun.

The basic strategy of the analyst, then, is to work through the text from beginning to end, using the mouse to select a single word or a stretch of text, to click on one of the cohesive marking buttons to bring up a list of option buttons, to click on some of the option buttons and/or references in the antecedent lists, and ending with an implicit or explicit confirmation (or cancellation) of the marking being constructed; seeing the marking introduced into the text window (and the antecedent list if appropriate); and repeat. There are subsidiary buttons, to provide a few other necessary requirements. A 'delete' button allows a complete cohesive marking to be removed (and then re-inserted in the correct form in the normal way, if required). Early designs of the XANADU program called for a range of editing buttons, to correct what were expected to be the common types of error, but we have not felt the need to incorporate these as yet, though an analysis of actual errors needing to be corrected in the text will allow us to make a rational choice of what to implement. A 'search' button allows simple or repeated searches on sequences of words forwards or backwards through the text.

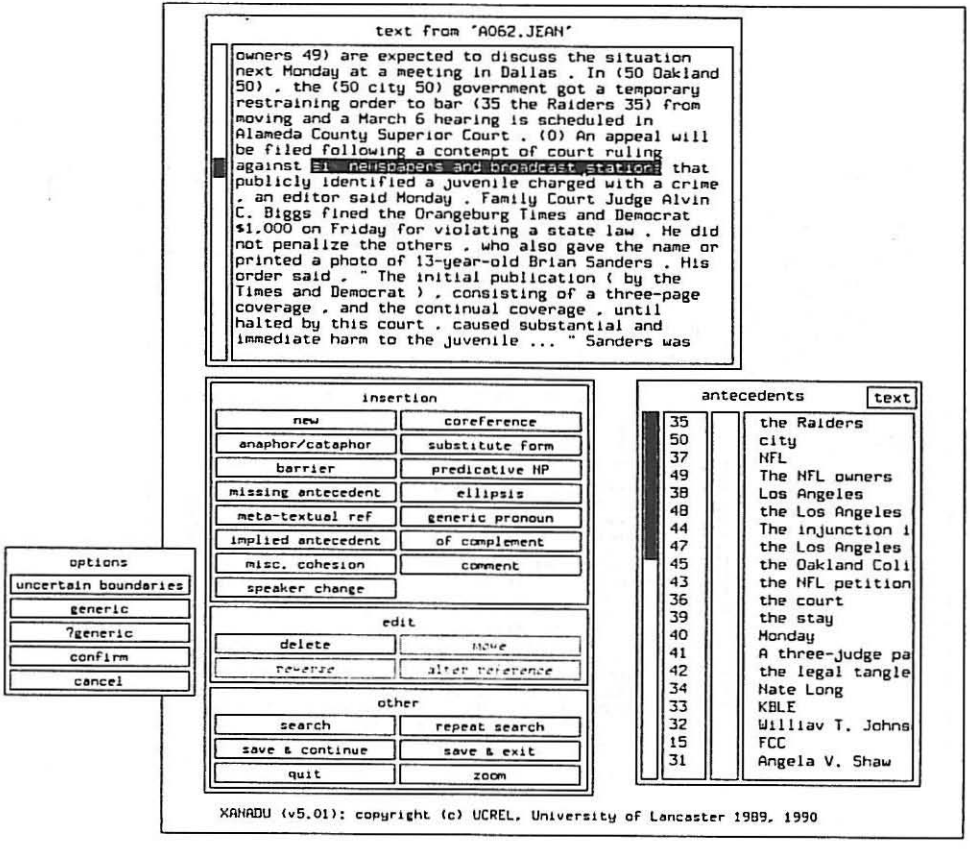

text from 'A062.JEAN'
owners 49) are expected to discuss the situation next Monday at a meeting in Dallas . In (50 Oakland 50) , the (50 city 50) government got a temporary restraining order to bar (35 the Raiders 35) from moving and a March 6 hearing is scheduled in Alameda County Superior Court . (0) An appeal will be filed following a contempt of court ruling against 31 newspapers and broadcast stations that publicly identified a juvenile charged with a crime , an editor said Monday . Family Court Judge Alvin C. Biggs fined the Orangeburg Times and Democrat $1,000 on Friday for violating a state law . He did not penalize the others , who also gave the name or printed a photo of 13-year-old Brian Sanders . His order said , " The initial publication (by the Times and Democrat) , consisting of a three-page coverage , and the continual coverage , until halted by this court , caused substantial and immediate harm to the juvenile ... " Sanders was
insertion
new
coreference
anaphor/cataphor
substitute form
barrier
predicative NP
missing antecedent
ellipsis
meta-textual ref
generic pronoun
implied antecedent
of complement
misc. cohesion
comment
speaker change
edit
delete
move
reverse
alter reference
other
search
repeat search
save & continue
save & exit
quit
zoom
antecedents
text
35 the Raiders
50 city
37 NFL
49 The NFL owners
38 Los Angeles
48 the Los Angeles
44 The injunction i
47 the Los Angeles
45 the Oakland Coli
43 the NFL petition
36 the court
39 the stay
40 Monday
41 A three-judge pa
42 the legal tangle
34 Nate Long
33 KBLE
32 William T. Johns
15 FCC
31 Angela V. Shaw
options
uncertain boundaries
generic
?generic
confirm
cancel
XANADU (v5.01): copyright (c) UCREL, University of Lancaster 1989, 1990

Figure 2

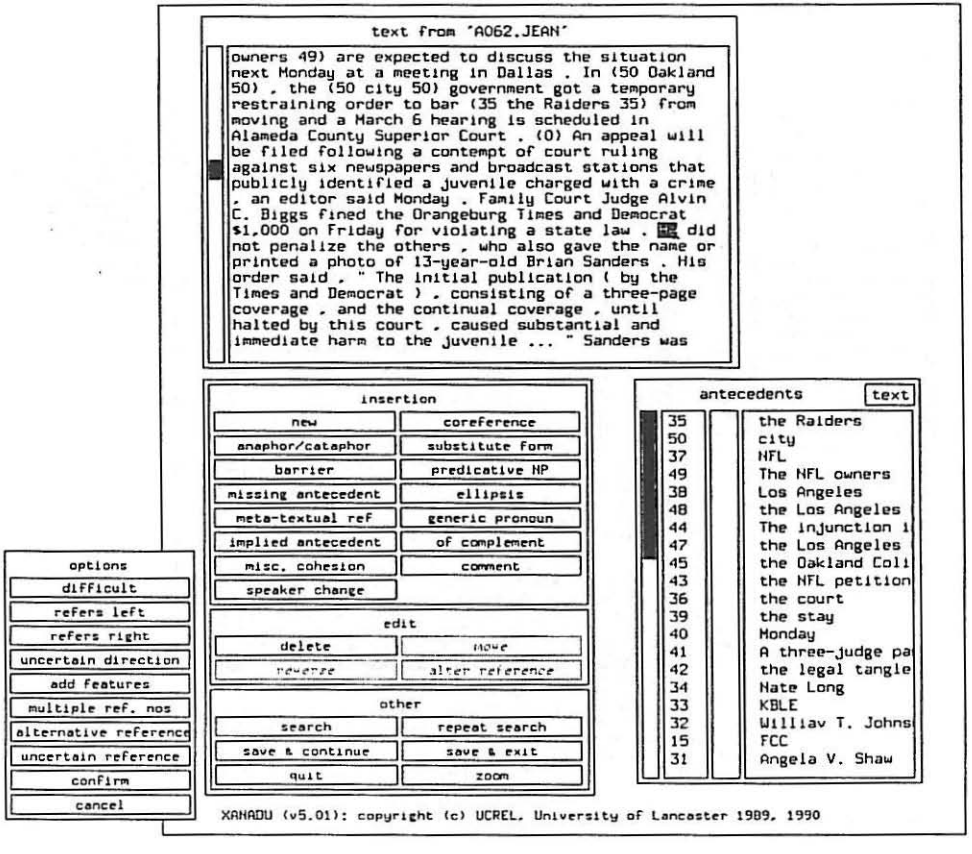

text from 'A062.JEAN'
owners 49) are expected to discuss the situation next Monday at a meeting in Dallas . In (50 Oakland 50) , the (50 city 50) government got a temporary restraining order to bar (35 the Raiders 35) from moving and a March 6 hearing is scheduled in Alameda County Superior Court . (0) An appeal will be filed following a contempt of court ruling against six newspapers and broadcast stations that publicly identified a juvenile charged with a crime , an editor said Monday . Family Court Judge Alvin C. Biggs fined the Orangeburg Times and Democrat $1,000 on Friday for violating a state law . He did not penalize the others , who also gave the name or printed a photo of 13-year-old Brian Sanders . His order said , " The initial publication (by the Times and Democrat) , consisting of a three-page coverage , and the continual coverage , until halted by this court , caused substantial and immediate harm to the juvenile ... " Sanders was
insertion
new
coreference
anaphor/cataphor
substitute form
barrier
predicative NP
missing antecedent
ellipsis
meta-textual ref
generic pronoun
implied antecedent
of complement
misc. cohesion
comment
speaker change
edit
delete
move
reverse
alter reference
other
search
repeat search
save & continue
save & exit
quit
zoom
options
difficult
refers left
refers right
uncertain direction
add features
multiple ref. nos
alternative reference
uncertain reference
confirm
cancel
antecedents
text
35 the Raiders
50 city
37 NFL
49 The NFL owners
38 Los Angeles
48 the Los Angeles
44 The injunction i
47 the Los Angeles
45 the Oakland Coli
43 the NFL petition
36 the court
39 the stay
40 Monday
41 A three-judge pa
42 the legal tangle
34 Nate Long
33 KBLE
32 William T. Johns
15 FCC
31 Angela V. Shaw
XANADU (v5.01): copyright (c) UCREL, University of Lancaster 1989, 1990

Figure 3

Buttons for 'quit' (that is, abandon the session, without retaining any of the markings inserted this time), 'save the current state of the text file and markings and continue the session' and 'save the current state of the text file and markings and terminate the session' provide overall control of the editing process.

More recently a special Unix shell script has been placed round the program to control attempted access by more than one analyst to the same file, to allow systematic renaming of the files being edited to show their status, and to allow hard-copy listings to be produced when required of the files being edited, with the syntactic marking retained or stripped out.

4. Further developments to the editing program

The previous section described the basic structure of the original design of the XANADU editor. Over the period of development of the annotation system there have been a number of changes in what is to be marked and how it is to be annotated. The analysts' experiences of the system in use have led to further changes. This section describes and discusses some of the consequential design changes to XANADU.

a. The expected use of the program was for the analyst to make essentially one pass through the text, marking the cohesive structures on the way through (with of course some local change of direction, to cope with cataphoric references and change of mind by the analyst). For this reason the antecedents already marked were displayed in a list ordered as follows; if several items had the same index number, then the textually latest one was chosen as the 'exemplar' of that index number; and these items were then displayed in textual order, with the textually latest at the top of the list. Thus when working at or near the end of the marked section of the text, the analyst would usually find any required antecedent among those most recently encountered, at or near the top of the list.

This organisation of the list works well in this first phase of editing. However, analysts turn out often to want to do a second pass through the text to make corrections, after thinking 'off-line' about particular annotation difficulties. Now the ordering of antecedents by their latest position within the text is often less helpful; the analyst scrolls to the appropriate part of the text window, and requires in the antecedent list those antecedents appropriate to that section. Various schemes were proposed, but no completely satisfactory way of meeting this requirement was agreed upon. Instead a rather simple scheme was implemented;

since the analyst usually knows the index number required (from a hard-copy listing of the text with cohesion marks included), the program now allows the antecedent list to be displayed either in textual order, as in the original system, or in numerical order of the index numbers, which allows simple scrolling to find the required antecedent for re-editing. The analyst can change back and forth between the two orderings as required by the editing task. An alternative strategy would, of course, have allowed an index number to be typed in instead of selected by the mouse (as was implemented in very early versions of the editing program), but we decided to stick with our principle of minimising the use of the keyboard.

b. The importance has already been mentioned in syntactic marking of allowing unlabelled brackets as a 'safety valve' for the analyst, so that a feature felt to be significant can be marked without a commitment to force it into the least unlikely of the categories provided. A similar safety valve was provided in the cohesive marking with the comment button. Here the analyst could click on a point in the text and then on the comment button, and insert a free-text note (which had to be typed at the keyboard) or any one of some ten to twelve pieces of 'canned' text attached to option buttons. The analyst was encouraged to use the latter mechanism if at all possible (perhaps supplemented with a free-text message), but the facility allowed significant problems or derogations from the guidelines to be marked (and these could then be searched for later, and edited into an alternative form if desired). In the most recent version of the program, the comment option buttons have been reduced to four, since another safety valve have been provided (see below): (i) to indicate a possible error in the syntactic marking; (ii) to indicate a point in the text requiring further checking by the referral system described in the next section; (iii) to indicate a case where the syntactic marking of the text, though correct, precludes the required cohesive marking – this is discussed further in (d) below; and (iv) to indicate an unusual or noteworthy cohesion feature.

Notation has now been provided for marking such marginal cohesive features as 'implied antecedents', and 'inferrable of-complements', and the alternative 'safety valve' of 'miscellaneous cohesion'. The guidelines for the first two of these have been drawn fairly tightly, and only a small number of these markings have occurred to date. It is planned that, when a substantial body of text has been marked for cohesion, all of these markings will be looked at again, with the possibility of modifying, eliminating or otherwise rationalising some or all of them.

An example of each of these markings (shown with all other cohesion notation suppressed, and taken from Fligelstone (1991), which contains a discussion of the guidelines for their use) is:

An implied antecedent:

He put on (3 his goggles 3) , fitted them tight , then tested <IMP=3(4 the vacuum 4) . [i.e. *the vacuum* in *his goggles*]

An inferrable *of*-complement:

He took the lead on the 65th lap of (3 the 80-lap race 3) and cruised to victory under the yellow caution flag after a car spun out on <OF=3(4 the 77th lap 4) . [i.e. *the 77th lap* of *the 80-lap race*]

Miscellaneous cohesion:

He will not discuss (3 prices 3) . But <MISC=3(4 the tab 4) comes to $4000 for a seven-point elk. [i.e. there is some cohesive link between *the tab* and *prices*]

In these three types of marking, there are potentially two links being introduced; one of the new cohesive relationships (indicated with the index number 3 in all the above examples), and a possible coreferential link from the item bracketed (here indicated with the index number 4). This results in two alternative orders in which the analyst might choose to introduce the links (if it is not done simultaneously). For example, in the last example, the analyst might originally bracket the words *the tab* as coreferential with some other item in the text, and later decide to link *the tab* with the word *prices* as a miscellaneous cohesion link; alternatively, the decision might bemade first to mark the miscellaneous cohesion between *prices* and *the tab*, with the coreferential link between the latter and another item inserted subsequently. In order to simplify the choices before the analyst, we decided that either the two links should be introduced together, or the first order must be used – thus in this example the miscellaneous cohesion link could be added to a pre-existing coreferential bracket round *the tab*. These two alternatives are indicated by the analyst clicking respectively on a stretch of words or a bracket. If a new bracket is being introduced, a new index number is also generated, in case coreferential items are later found.

c. Most of the cohesion markings inserted by the analyst are attached to a sequence of one or more orthographic units (words and possibly punctuation marks). The marking of ellipsis is special in that it is attached, not to a series of words, but to a point in the text from which it is to be understood that material has been ellipted. Thus, in:

John (11 was eating 11) an ice cream, and Mary <ELIP=11 a bun.

the notation indicates that the sentence is to be understood as having the antecedent *was eating* ellipted from the position marked. It should be remarked that the guidelines for ellipsis require it to be marked only in cases where it can be recovered from the text, although small morphological modifications to the antecedent are allowed to adjust it to the context of the ellipsis site.

There is sometimes a problem with the insertion of an ellipsis marking, since the text being marked for cohesion has already been marked for the major syntactic structures of the sentence. Since the syntactic markings are not visible during the editing process, a single position in the visible text could correspond to several possible positions in the syntactically marked text. Consider the sentence:

The boy (169 sat 169) in the back, and the girl <ELIP=169 in front.

Here we have omitted the syntactic marking of the sentence, just as it would be displayed in the text window. The ellipsis marking for the ellipted word *sat* lies between the words *girl* and *in*. However the syntactic labels N] and [P also lie between *girl* and *in* (ie closing the noun phrase *the girl* and opening the prepositional phrase *in front* respectively). It is not possible for the editing program to decide which is the correct position among these syntactic brackets. In this type of situation, where the program detects alternative positions for the ellipsis marker, it displays a window showing an appropriate portion of the text including the syntactic markings, allowing the analyst to click the mouse on the appropriate position. An example of such a window is shown in Figure 4.

d. Most of the sequences of orthographic units marked with the cohesion annotation are grammatical constituents, and an attempt to mark a sequence of units which violates the hierarchical structuring of the constituents will usually be an error. For this reason, the program checks any sequence of units selected for marking against the hidden syntactic marking (and also against the other cohesion markings in the vicinity), to ensure that the markings form a hierarchical tree. When this check fails, the program displays an error message, and a window appears showing the words and the detailed syntactic marking in the area of the attempted insertion. There is also a 'zoom' button, a recently installed feature which allows the detailed syntactic structure of any selected sentence in the text window to be displayed in a pop-up window, to allow inspection by the analyst in the few cases where this is important.

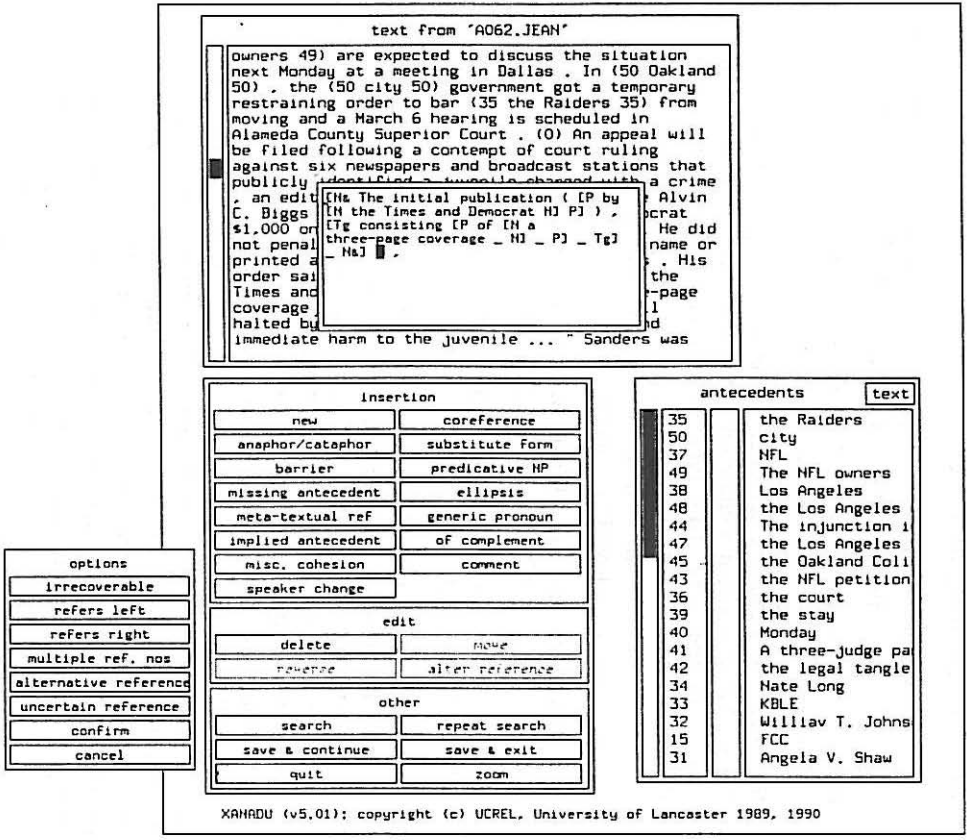

text from 'A062.JEAN'
owners 49) are expected to discuss the situation next Monday at a meeting in Dallas . In (50 Oakland 50) , the (50 city 50) government got a temporary restraining order to bar (35 the Raiders 35) from moving and a March 6 hearing is scheduled in Alameda County Superior Court . (0) An appeal will be filed following a contempt of court ruling against six newspapers and broadcast stations that
[N& The initial publication ([P by [N the Times and Democrat N] P]) , [Tg consisting [P of [N a three-page coverage _ N] _ P] _ Tg] _ N&] ,
immediate harm to the juvenile ... " Sanders was
insertion
new
coreference
anaphor/cataphor
substitute form
barrier
predicative NP
missing antecedent
ellipsis
meta-textual ref
generic pronoun
implied antecedent
of complement
misc. cohesion
comment
speaker change
edit
delete
move
reverse
alter reference
other
search
repeat search
save & continue
save & exit
quit
zoom
options
irrecoverable
refers left
refers right
multiple ref. nos
alternative reference
uncertain reference
confirm
cancel
antecedents
text
35 the Raiders
50 city
37 NFL
49 The NFL owners
38 Los Angeles
48 the Los Angeles
44 The injunction i
47 the Los Angeles
45 the Oakland Coli
43 the NFL petition
36 the court
39 the stay
40 Monday
41 A three-judge pa
42 the legal tangle
34 Nate Long
33 KBLE
32 William T. Johns
15 FCC
31 Angela V. Shaw
XANADU (v5.01): copyright (c) UCREL, University of Lancaster 1989, 1990

Figure 4

There have turned out to be two main cases where it is important to be able to mark a stretch of text in a way which violates the hierarchical structure of the syntactic markings:

(i) an antecedent covering a sequence of sentences together with a partial sentence, typically in reported direct speak, for example:

John said , ' (12 This is what I propose. We go to London tomorrow. 12) ' He told me <REF=12 this at lunch.

In the simple sentence-by-sentence skeleton parsing used by the UCREL analysts, the antecedent numbered 12 covers half of one sentence and the whole of the next. This type of structure occurs sufficiently often in the texts being annotated that the program has been designed to test for this type of structure when checking the match with the syntactic bracketing, and to allow it to stand.

(ii) an antecedent for an ellipsis which does not respect constituent boundaries, for example sentences A038/51-2:

[N (13 The circus N] [V is 13) [J intricately accurate J] V] .
<ELIP=13 So intricate and accurate that ...

Here there is ellipted material at the beginning of the second sentence, and we would like to indicate that this is the words *the circus is*. As can be seen from the partial syntactic marking of the first sentence, the marking of the ellipted material would intersect a syntactic marking.

The UCREL team has mixed views as to how to deal with this. The early versions of the XANADU software checked the consistency of each cohesion marking against the syntactic bracketing (and other cohesion marking), as described above. A later version of the software queried any cohesion marking which was inconsistent with the syntactic bracketing, but allowed the analyst to override the check and insert the marking. The idea was that, at a later stage the program could check that any such 'ungrammatical' antecedents were referred to only by ellipsis markings. In later versions of the program we have removed this feature, and the program does not allow an 'ungrammatical' marking to be inserted (except the special case of (i) above), on the grounds that we would be losing a valuable check for what is usually an error. Situations where such markings are required have turned out to be very rare, and where necessary are marked with a suitable cohesion comment. This may not be our final word in this area; if we do allow such things to be marked, it must be done in such a way that the annotation can

be made to conform to other text mark-up systems, such as TEI, which have special ways of indicating non-hierarchical markings.

5. The annotation process as a whole

The XANADU editing software is central to the process of marking cohesion features in the text, but there is a sequence of subsidiary processes in organising a procedure for turning out high-quality annotated text. The main steps of the procedure are as follows:

a. Since we are inserting cohesion markings only on texts which have been syntactically marked, an earlier stage of the process involves the selection and preparation of the texts, their automatic word-tagging followed by manual post-editing of the word-tags, and then the manual skeleton parsing process is carried out. This process is described in detail elsewhere. Suitable texts resulting from this process are selected and placed in the cohesion text directory.

b. An analyst selects a text from this directory and edits it using the XANADU editor. At this point the text is automatically logged out to this analyst, so that it is not available for editing by anyone else. A Unix shell script has been written to protect the analysts from the full force of the Unix shell, and to allow the files being edited to be systematically and automatically renamed to show who is the analyst responsible for that block and what stage the analysis has reached. This script also allows the analysts to obtain hard-copy listings of partially-analysed texts, with or without the syntactic marking displayed.

c. The analyst makes a first pass over the whole text, marking the cohesive structures found there. There will usually be difficulties in deciding how to mark certain features. The analyst can re-edit the partially-marked text in due course after having made decisions on the outstanding issues, perhaps by consulting other colleagues or hard-copy listings of the text.

d. In the early days of this project all of the output from the analysts was checked by one person (Steve Fligelstone) for adherence to the guidelines he was developing. This ensured the adequacy of the guidelines, and also helped with the training of the analysts. Now that the cohesion marking is moving into production it is planned that quality control will be assured by similar procedures to those used for the skeleton parsing project. Thus a proportion of the blocks of text (5-10%) will be marked by one analyst and cross-checked by a second analyst, and

differences resolved by discussion. There will be a team of checkers in the Linguistics Department of the University of Lancaster who will examine a smaller proportion of the output from the analysts – it is planned that they will look at 1-5% of the texts, examine all the cohesion comments, and act as referees in case of disagreement between analysts on how a passage should be marked. Although any errors shown up by the marking are corrected, the main aim is to check on trends in the annotation, for feedback directly to the analysts and via the manual of guidelines, a document of some 100 pages. This manual went through rapid development in the early stages of the project, and has now settled done with occasional additions.

e. In the planning of the project it was expected that there would be a post-processing program to check and augment the human analysis. A number of small 'filter' programs are at present run over the annotated text during the checking stage described above, to search for certain common problem areas. It is planned that a further program will shortly be implemented to carry out all these checks automatically, tidy up the annotation (by eliminating unused antecedent markings and resequencing the antecedent numbers, for example), and insert any further markings which can be done automatically. An example of the latter is identical naming expressions identified by a set of rules known to both the analysts and this program, allowing the analyst to omit the mark-up in these cases; it would be possible to present doubtful cases to the analyst for further consideration, but our aim it to minimise further manual intervention at this stage.

6. Results and conclusions

After a period of development, the XANADU program is in daily use by a team of three analysts producing texts with cohesion markings. It is difficult to give figures for the speed of editing, since the density of cohesion marking of texts varies much more than that of the syntactic marking of sentences, but a text of some two thousand words would receive a first pass of annotation in about half a day. There would usually be some re-editing of this text after consideration of any problems, or as a result of the quality-checking procedure. This is of course a peak rate, as the analysts are also engaged in syntactic marking of texts.

Most of the cohesion marking has been done on a corpus of Associated Press news stories which was skeleton parsed by the UCREL team over the period 1988 to 1989. Two other corpora which have been syntactically

marked have proved less amenable to cohesion marking. A corpus of IBM Computer Manuals is currently being skeleton parsed; this was constructed by selecting sentences controlled for vocabulary, and therefore does not consist of consecutive sentences. A corpus from (the English recension of) a portion of the Canadian Hansard, also parsed over the period 1988 to 1989, is constructed from consecutive sentences but has also turned out to be difficult to annotate; perhaps this is a comment on the speech patterns of politicians. We expect to do some further cohesion marking on texts from a smaller corpus we are building, a collection of twenty-five million words of middlebrow British and American English.

To date (Summer 1992) a total of some 5442 sentences (50 texts) from the AP corpus has been marked for cohesion. This includes texts marked in the early stages of development of the annotation scheme, since they have been re-edited in accordance with the current guidelines. A preliminary census has been made of types of cohesion markings found in these blocks, with a view to checking the design assumptions of the XANADU program. Apart from the cohesion marking, these texts contained 110822 words and 14239 punctuation marks, and 143044 syntactic brackets (where a pair of labels such as '[N ... N]' is counted once). There were 20553 cohesion markings (where a bracketed item such as '(77 ... 77)' is counted once). These cohesion markings were broken down as follows:

Number of antecedents and coreferential items: 12409 (a typical set of non-proform coreferential items would have two to three such items linked to the antecedent; note that the marking of an antecedent is distinguished from that for a non-proform coreferential item not by form, but only by position).

Number of referential proforms: 4151 (3998 anaphoric, 134 cataphoric, 19 uncertain)

Number of non-referential pronouns: 191 (here the pronoun is marked as having certain features, such as singular or plural *you*, or to link it with a following co-referring pronoun, but it does not refer to a non-proform item in the text)

Number of substitute forms: 425 (7 cataphoric)
Number of ellipsis markings: 97 (1 cataphoric)
Number of meta-textual references: 5 (all cataphoric)
Number of markings of copular relationships: 1796 (362 cataphoric)
Number of implied antecedent items: 222 (10 cataphoric)

Number of inferrable *of*-complementations: 226 (19 cataphoric)
Number of miscellaneous cohesion marks: 609 (17 cataphoric)
Number of cohesion barriers: 406

Number of cohesion comments: 16 (of these six called attention to some issue over the interaction between the syntax and the cohesion marking, and ten called attention to some noteworthy curiosity of the cohesive structure of the text – it should be borne in mind that these are the residual comments after the text has been right through the checking procedure).

The discussion in Section 2 indicates that the editor was designed on the assumption that proforms more than one word long would be rare. This has turned out to be the case – there were 4836 proforms which were one word long, and 33 others. These others were all two words long; two-thirds of them were substitute forms, mostly of the form *do it*, *does so*, etc; most of the remaining third were referential *each other*.

Similarly the editor is designed on the assumption that references to a single antecedent would be the most common, and this has also turned out to be true. There were 20746 references to single items, 323 multiple references, and 33 alternative references. There were no examples among these texts of mixed multiple and alternative references, of the '>REF=1,5/6,?22' form discussed in Section 2.

Other points worthy of note in the above preliminary figures are the low ratio of cataphoric to anaphoric cohesion, and the large number of 'miscellaneous cohesion' items – this is perhaps inevitable in an annotation with such a name, but invites further analysis of the examples found, leading perhaps to sharper guidelines for the analysts. Another feature of the initial analysis was the number of antecedents marked by the analysts as possible sites for a cohesive link, but then not in fact used; the counts of these links have been subtracted from the above figures. These would normally be removed automatically by the post-processing program mentioned above.

The XANADU cohesion editor is now in a stable condition, and it is not expected that there will be significant further developments in its use for cohesion marking. In this area the next software development is likely to be of a program (or suite of programs) for automatic checking of as much as possible of the cohesion-marked texts, and for augmentation and rationalisation of manually-analysed texts where this can be done automatically.

However, we believe that further developments are possible in the use of the general XANADU editor framework in other areas where

texts need to be annotated in complex ways. We are investigating the use of a version of the program in the manual indication of prosody on text derived from speech, and as an alternative to the special-purpose editor we use for skeleton parsing. We expect this to lead us to develop a general-purpose version of the XANADU software in which the particular annotation symbols, and the rules under which they are allowed to interact, are specified by tables within the program, rather than being hard-wired. Other areas where we believe that such a general-purpose text annotator would be useful are are in pragmatic and discourse annotation.

Note

1. Thanks are due to Geoffrey Leech and Steve Fligelstone of UCREL and Ezra Black of IBM for the development of the notation; to Steve Fligelstone for early discussions of the form a task-oriented editor should take; to Jean Forrest, Liz Eyes and Simon Botley for using and critiquing the notation and the XANADU editor; and to the IBM Thomas J. Watson Research Center, Yorktown Heights, and IBM Scientific Centre, Winchester, for funding.

References

Fligelstone, S.D. 1991. A description of the conventions used in the Lancaster anaphoric treebank scheme. Internal UCREL report, available from: UCREL, University of Lancaster, Bailrigg, Lancaster LA1 4YR, U.K.

Fligelstone, S.D. 1992. Developing a scheme for annotating text to show anaphoric relations. In *Proceedings of the 11th International Conference on English Language Research on Computer Corpora*, ed. by G. Leitner. Berlin: Mouton de Gruyter.

Halliday, M.A.K. and R. Hasan. 1976. *Cohesion in English.* London: Longman.

Leech, G.N. and R.G. Garside. 1991. Running a grammar factory: The production of syntactically analysed corpora or 'treebanks'. In *English computer corpora: Selected papers and research guide*, ed. by S. Johansson and A. Stenström. 15-32. Berlin: Mouton de Gruyter.

Sperberg-McQueen, C.M. and L. Burnard (eds). 1990. *Guidelines for the encoding and interchange of machine-readable texts.* Draft Version 1.0. Chicago and Oxford: ACH, ACL, ALLC.

Quantitative comparison of modals in the Brown and the LOB corpora[1]

Junsaku Nakamura
University of Tokushima

1. Introduction

The laborious task of comparing word frequencies in British and American English was undertaken by Hofland and Johansson (1982), using the LOB and the Brown corpora. Not only did it provide an alphabetical list of all the words which occur ten times or more and are distributed in at least five text samples in one of the two corpora accompanied by a 'difference coefficient', but it also gave some interesting observations concerning differences between the two corpora in terms of spellings, word forms, uses of auxiliary verbs, and uses of words in some semantic groups.

The treatment of modal auxiliaries, however, was quite cursory, only comparing the total frequencies in the two corpora and drawing the conclusion that the use of modals is quite similar except for the possible cases of *ought*, *shall* and *should*, the difference coefficients of which indicate that they are a little overrepresented in the LOB Corpus. But as Hofland and Johansson (1982: 36) correctly observed, 'the frequency of the modals varies considerably in the text categories and should therefore be subjected to a more detailed examination'.

Following up this remark of theirs, the present study first explores the differences in the use of modals across various genres in the two corpora and then compares the corpora from the viewpoint of the use of modals. This comparison is performed purely on a statistical or quantitative basis, by means of a statistical technique called Hayashi's Quantification Method Type III, which has been successfully used in various works of the present author as shown in the list of references.

2. *Data and method*

2.1. Data

The ICAME CD-ROM (MS-DOS version) containing the Brown, Helsinki, Kolhapur, London-Lund and LOB corpora was used for obtaining the data for the following discussion either by means of the WordCruncher and the TACT programmes, which are provided together with the CD-ROM. The simple method consisted of extracting the references of the word or words in question from the word selection list, sorting them in case several forms are involved, and counting the number of occurrences according to the first alphabetical letter of the reference indicating the genre. This was adequate for counting the frequencies of modals across genres in the LOB Corpus, in which all the words are tagged and contractions are treated separately.

The matter was not so simple in case of the Brown Corpus since unfortunately the tagged version is not available on the CD-ROM. So counting was first done by computer in terms of graphic words. Disambiguation, if necessary, was performed manually by looking through the concordance lines, and finally the figures in the frequency tables were adjusted accordingly.

Ten modals treated in Hofland and Johansson (1982), i.e. *can, could, may, might, must, ought, shall, should, will,* and *would*, plus *need, dare,* and *used* were included in the analysis. The first two items added in this study, i.e. *need* and *dare*, were counted as modals only if tagged as modals in the LOB Corpus, and the same criteria were used for disambiguating the Brown Corpus. The item *used* is treated as a verb taking a *to*-infinitive as its complement in the LOB Corpus, but often it is treated as a semi-modal like *need* and *dare*. In this study *used* in association with *to* indicating some regular activity or state in the past was treated as modal.

Since the number of occurrences of *dare, need*, and *used* is very limited as will be shown below, the results of the present study would not be influenced very much if they were included in or excluded from the list. (So anyone who is unhappy with the inclusion of these items can ignore the references to them in the following discussion.)

Spelling variants were included in the frequency counts, referring to the lemmatized alphabetical list of Francis and Kučera (1982). In the case of *dare*, both present and past forms were counted together because the number of occurrences was very small.

Table 1, which is actually composed of two tables, one for the Brown Corpus and the other for the LOB Corpus, compressed into one, shows the results of this count and provides the basic data for the present study. The raw frequency data given in Table 1, however, are from genres of different sizes, whereas the ways modals are used in each genre should be independent of item size. Therefore, the frequency figures given in Table 1 were next adjusted to the average genre size of each corpus.[2] This adjusted frequency table for each corpus was then fed into the statistical procedure called Hayashi's Quantification Method Type III for the analyses of the individual corpora which will be discussed in the first part of Section 3. Later, the frequencies of the whole table were adjusted to the average genre size covering both the Brown and the LOB corpora and processed again by Hayashi's Quantification Method Type III for the analysis of the combined corpus discussed in the last part of Section 3.

2.2. Method

2.2.1. Hayashi's Quantification Method Type III

Generally speaking, Hayashi's Quantification Method Type III is used for quantifying qualitative or attributive categories and samples simultaneously. The distinctive feature of this method is that it can classify or quantify both categories and samples only by looking at qualitative dichotomic response patterns, i.e. depending upon whether each sample reacts positively or negatively to several categories, without requiring any external criterion. Given a data matrix, the size of which is the number of samples by the number of categories, the basic principle is to rearrange the rows and the columns so that the positive responses converge around the diagonal. Consequently, the categories placed close to one another and the samples close to one another are considered to be qualitatively similar and those located at a distance from one another are said to be qualitatively different.

Hayashi's Quantification Method Type III performs this task of rearrangement of the data matrix, not literally but numerically, giving the categories close to one another numerical values (called category weights) which are close to one another and giving the samples close to one another numerical values (called sample scores) close to one another. Using the quantities x_i and y_j which represent the quantities given to samples and categories respectively, the procedure of this technique is to determine the numerical values of x_i and y_j which maximize the correlation coefficient between x and y.[3] The algorithm extended to

Table 1. Frequencies of Modals in the Brown and the LOB Corpora.

Corpus	Genre	CAN	CLD	MAY	MGT	MST	OGT	SHL	SHD	WLL	WLD	NED	DRE	USD	TOTAL	SIZE
	A	124	124	66	38	52	1	5	64	427	253	1	0	2	1157	88690
	B	166	168	74	38	55	5	19	93	245	193	1	0	2	1059	54505
	C	52	53	45	26	18	3	2	18	61	49	3	0	1	331	34346
	D	106	106	80	12	54	4	23	45	67	71	4	2	0	574	34590
	E	303	306	130	22	82	0	5	74	310	87	0	1	3	1323	72590
	F	206	207	165	47	96	8	12	78	182	203	5	1	8	1218	97223
	G	313	313	214	113	170	6	35	109	242	416	12	0	10	1953	152064
BROWN	H	133	133	155	13	102	1	99	113	242	120	1	0	0	1112	62477
	J	425	212	323	128	203	8	42	179	337	322	10	0	3	2192	162211
	K	54	24	8	42	57	4	3	38	98	333	0	0	6	667	58380
	L	70	34	13	57	33	5	4	30	117	280	0	0	2	645	48204
	M	24	8	4	12	8	1	3	4	28	80	0	0	0	172	12042
	N	73	27	6	59	28	7	11	20	150	249	0	0	4	634	58416
	P	115	46	11	51	50	14	4	43	157	327	1	0	8	827	58625
	R	29	15	8	8	9	3	2	7	28	64	0	0	2	175	18277
SUBTOTAL		2193	1776	1302	666	1017	70	269	915	2691	3047	38	4	51	14039	1013644
	A	123	107	71	44	63	3	14	120	312	254	7	0	4	1122	89138
	B	180	92	93	41	89	7	9	146	239	195	4	0	3	1098	54447
	C	68	50	19	8	33	2	4	19	73	39	0	0	2	317	34321
	D	103	36	65	18	50	8	25	41	100	101	6	0	5	558	34387
	E	266	77	117	43	84	2	15	146	316	154	8	0	2	1230	76913
	F	186	101	128	54	81	3	8	98	198	159	11	1	17	1045	89090
	G	290	257	190	129	156	25	26	187	200	409	9	0	16	1894	155336
LOB	H	110	40	162	55	75	5	95	126	157	147	4	0	2	978	60761
	J	387	144	385	115	206	13	60	204	297	315	9	0	0	2135	161900
	K	84	190	21	60	87	4	28	64	96	288	3	2	11	938	59204
	L	74	183	15	68	48	11	11	35	49	177	4	1	5	681	49145
	M	17	25	15	15	18	0	1	11	20	35	0	1	1	159	12119
	N	106	203	17	59	57	11	15	40	82	173	2	2	5	772	59391
	P	121	200	27	53	85	8	41	49	130	289	6	4	5	1018	59382
	R	32	36	13	17	15	1	3	15	47	64	3	0	6	252	18203
SUBTOTAL		2147	1741	1338	779	1147	103	355	1301	2316	2799	76	11	84	14197	1013737
TOTAL		4340	3517	2640	1445	2164	173	624	2216	5007	5846	114	15	135	28236	2027381

deal with ordinary cross tables containing frequency figures was used to process the adjusted frequency tables mentioned above with the thirteen modals treated as categories and the fifteen genres as samples.

If there are n categories and more than n samples, Hayashi's Quantification Method Type III can produce n-1 sets, or "axes", to use a statistical term, of category weights and sample scores; that is, in the present case, there are twelve ways to assign numerical values to modals and genres, each of which yields a correlation coefficient lower than that of the set produced in the previous stage. But only the first three sets were used here. As shown in Table 2, the cumulative proportion accounted for by these three axes amounts to about 90.1% in case of the Brown Corpus, which means that less than 10% of the information contained in Table 1 was left unaccounted for. The same figure for the LOB Corpus is not so high as for the Brown Corpus (about 83.9%), but the rest of the axes were nevertheless discarded, since one can include only up to 3 axes in the figures which visualize the relative positions of modals or genres according to the quantities given to them.

Table 2. Correlationship Coefficients, Proportion Accounted for, and Cumulative Proportion Accounted For for Each Axis.

Corpus	Axis	Correlationship Coefficients	Proportion Accounted for	Cumulative Proportion Accounted For
Brown	1	0.3958423D+00	0.6555276D+02	0.6555276D+02
	2	0.1802609D+00	0.1359407D+02	0.7914683D+02
	3	0.1617245D+00	0.1094203D+02	0.9008887D+02
LOB	1	0.3091174D+00	0.5668550D+02	0.5668550D+02
	2	0.1773279D+00	0.1865431D+02	0.7533981D+02
	3	0.1200938D+00	0.8555912D+01	0.8389572D+02

The three sets of category weights (i.e. quantities given to modals) and sample scores (i.e. quantities given to genres) calculated for producing the above correlation coefficients were then normalized, with the means equal to 0, and the variances equal to 1.0. Normalized category weights and sample scores thus calculated are given in Table 3 and Table 4.

2.2.2. Plotting the quantities in a three-dimensional space

Although these numerical values indicate the relative positions of each modal or each genre along three axes, it is not so easy to grasp the overall picture. Mere numerical figures are always difficult to process.

Table 3. Quantities Given to Genres.

Corpus	Genre	Axis 1	Axis 2	Axis 3
BROWN	A	0.8644938D-01	-.1118521D+01	-.2416035D+01
	B	-.3690348D+00	-.5404565D+00	-.7454465D+00
	C	-.5626481D+00	-.7069115D+00	0.8569155D+00
	D	-.1125388D+01	0.6554862D+00	0.1117931D+01
	E	-.1153327D+01	-.1809942D+01	-.2578692D+00
	F	-.6667661D+00	-.3639210D+00	0.1029184D+01
	G	-.3535292D+00	0.2538856D-01	0.1242463D+01
	H	-.1125394D+01	0.2402646D+01	-.1444348D+01
	J	-.6452923D+00	0.4024467D+00	0.8788340D+00
	K	0.1585614D+01	0.6219538D+00	0.4251110D+00
	L	0.1335521D+01	0.6902562D-01	0.9648778D-01
	M	0.1361809D+01	0.3751446D+00	0.2282633D+00
	N	0.1270818D+01	0.6983861D-02	-.6607772D+00
	P	0.1143489D+01	-.1505324D+00	0.1051702D+00
	R	0.7775732D+00	-.1351424D+00	0.5380980D+00
LOB	A	0.3818902D+00	-.1184028D+01	0.1894569D+01
	B	0.6783344D+00	-.9293684D+00	0.1882208D+00
	C	0.2131540D+00	-.1377625D+01	-.5534364D+00
	D	0.7885892D+00	0.4119688D+00	0.1343637D+00
	E	0.1160080D+01	-.1325366D+01	0.6923686D-02
	F	0.6408258D+00	-.5846880D+00	-.9243607D+00
	G	-.1114538D+00	0.2986858D+00	-.1164863D+01
	H	0.1359840D+01	0.2517254D+01	0.1065685D+01
	J	0.1098158D+01	0.8108970D+00	-.1829384D+01
	K	-.1276520D+01	0.4447176D+00	0.9009403D+00
	L	-.1751090D+01	0.3446721D+00	-.7402488D+00
	M	-.4980389D+00	0.2739851D+00	-.1342692D+01
	N	-.1433072D+01	-.3918005D-01	-.6171378D+00
	P	-.1052996D+01	0.4218404D+00	0.1028152D+01
	R	-.4787470D+00	-.6684050D+00	0.8082985D+00

Therefore, the values in Table 3 and Table 4 are next plotted in Figure 1 through Figure 4, showing the relative position of each item in a three-dimensional space. In these figures, the genres sharing similar tendencies in the use of modals or the modals showing a similar distribution across genres assume the same position or are placed close to one another. If a particular genre were found to exhibit an even distribution of modals, it would be placed at the origin of the coordinates, showing that there is no preference whatsoever in the use of modals in that genre. In the same way, if a particular modal were used evenly across genres, it would also be placed in the origin of the coordinates, showing that the use of this modal is not influenced by genres at all. In general, if an item is located very close to the origin, then it is supposed to be rather neutral, whereas those located far away from the origin can be said to be unique in one way or another.

Table 4. Quantities Given to Modals.

Corpus	Modal	Axis 1	Axis 2	Axis 3
BROWN	CAN	-.4165721D+00	-.6123355D+00	0.6298268D+00
	COULD	-.1117624D+01	-.1051435D+01	0.3547252D+00
	MAY	-.1500111D+01	0.7049650D+00	0.9132860D+00
	MIGHT	0.1008490D+01	-.1847013D+00	0.9451326D+00
	MUST	-.3695249D+00	0.8806653D+00	0.6879512D+00
	OUGHT	0.1157511D+01	0.2120927D-01	0.1725852D+01
	SHALL	-.1319560D+01	0.5353011D+01	-.1850965D+01
	SHOULD	-.6234017D+00	0.7979112D+00	-.4026735D+00
	WILL	-.1718171D-01	-.6304738D+00	-.1761899D+01
	WOULD	0.1428600D+01	0.3183936D+00	0.2833557D+00
	NEED	-.1576262D+01	0.4304760D+00	0.4688756D+01
	DARE	-.2709357D+01	0.6248808D+00	0.5412382D+01
	USED	0.1274516D+01	-.9505088D+00	0.1562803D+01
LOB	CAN	0.5281552D+00	-.6879480D+00	-.8777260D+00
	COULD	-.1671419D+01	-.8779787D-01	-.4822368D+00
	MAY	0.1665452D+01	0.1267627D+01	-.1507674D+01
	MIGHT	-.8112135D+00	0.7464739D+00	-.1072108D+01
	MUST	-.1100047D-01	0.2560502D+00	-.7424606D+00
	OUGHT	-.8222917D+00	0.9099377D+00	-.1257105D+01
	SHALL	0.8111442D+00	0.4412333D+01	0.2924380D+01
	SHOULD	0.8929053D+00	-.5211835D-01	0.3337108D+00
	WILL	0.7751807D+00	-.1221471D+01	0.1067458D+01
	WOULD	-.7555002D+00	0.1904645D+00	0.7658680D+00
	NEED	0.3347839D+00	-.4337187D+00	0.1253263D+01
	DARE	-.3013300D+01	0.1476729D+01	-.1917553D+01
	USED	-.9681253D+00	-.7580426D+00	0.5675116D+00

These figures contain another piece of information obtained from the adjusted frequency table, i.e. the relative size of the item in question indicated by the size of a round figure placed at the tip of the vertical line. (These round figures are to be considered not as circles but rather as balls since these figures are supposed to be three-dimensional.) For example, Figure 1 shows that Genres B, E and H use a large number of modals in contrast to Genre C with the smallest number.

Now that the numerical values are all transformed into a point in a three-dimensional space, it becomes possible to grasp visually and easily the relationships of items to one another.

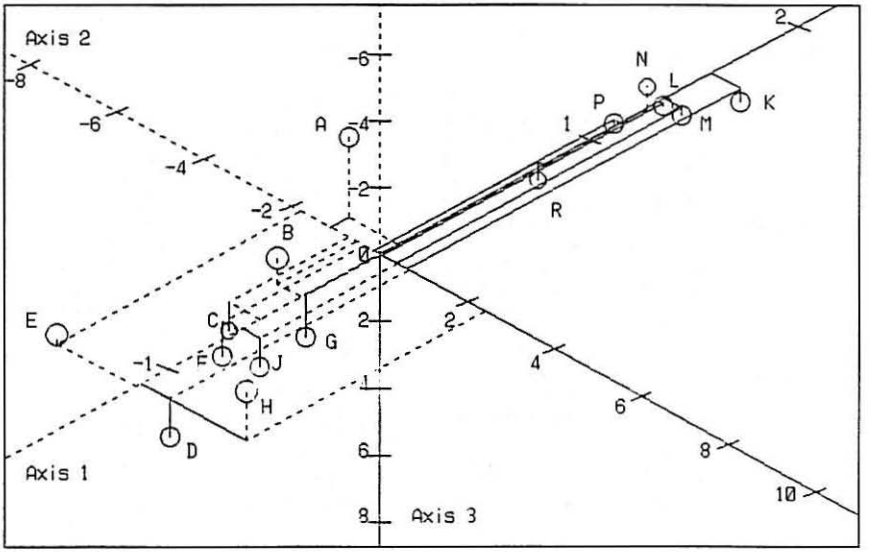


Fig. 1. Three-Dimensional Distribution of Genres in the Brown Corpus.

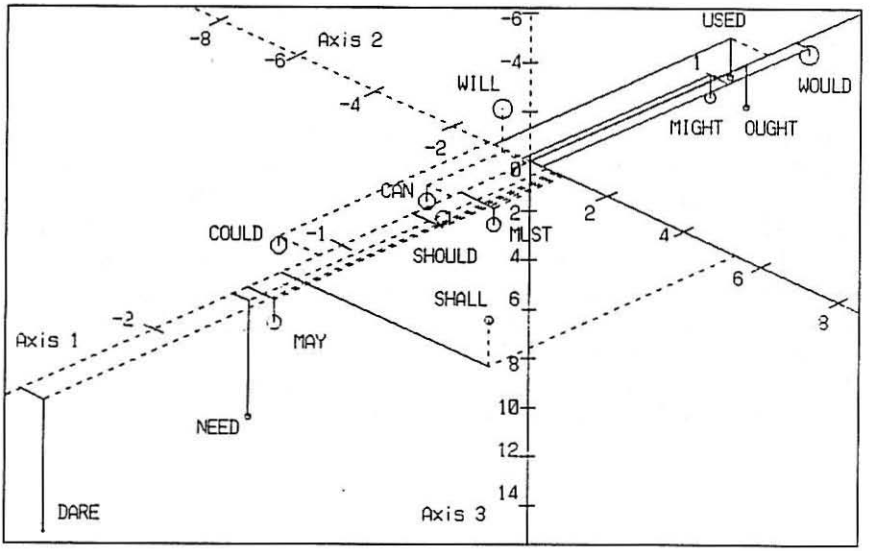


Fig. 2. Three-Dimensional Distribution of Modals in the Brown Corpus.

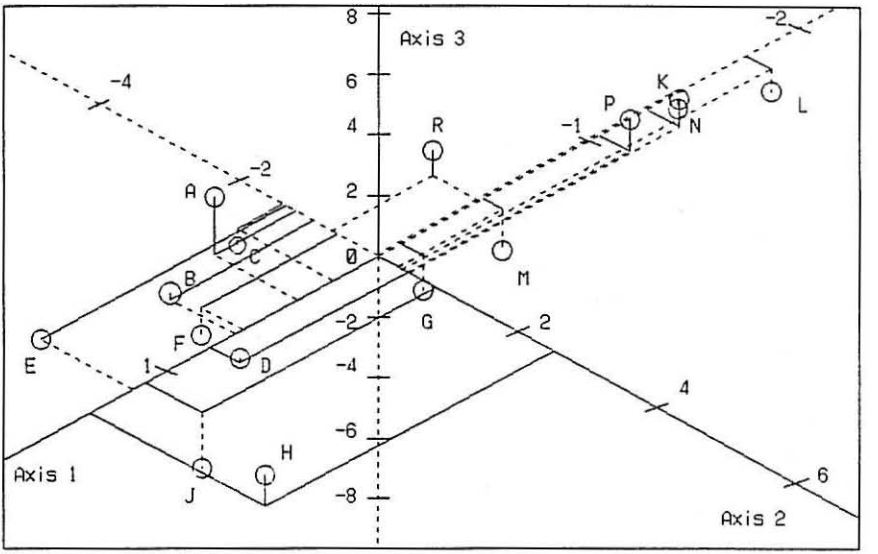


Fig. 3. Three-Dimensional Distribution of Genres in the LOB Corpus.

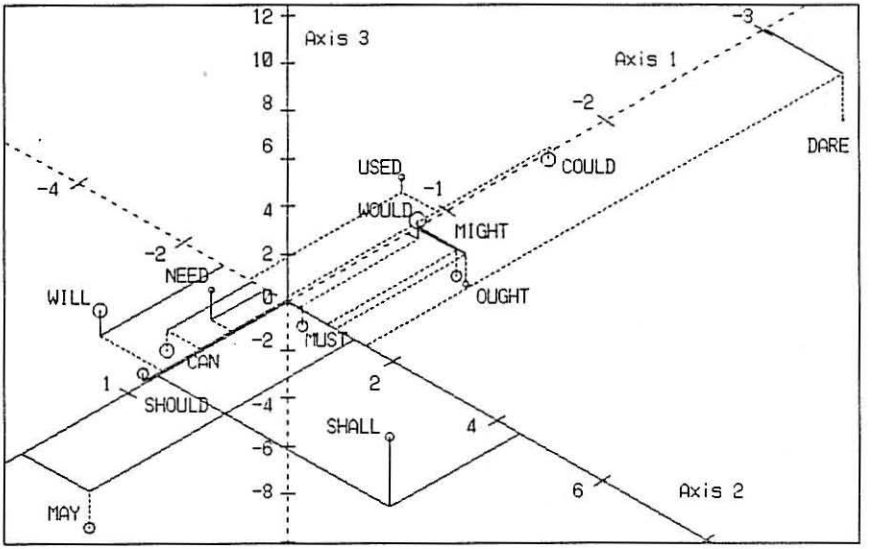


Fig. 4. Three-Dimensional Distribution of Modals in the LOB Corpus.

3. *Results*

3.1. The Brown Corpus

As is shown in Figure 1,[4] the genres of informative and imaginative prose in the Brown Corpus are rather neatly separated, except for Genre A (Press: reportage). All the genres of imaginative prose are located in the negative range of Axis 1, whereas most of the genres belonging to informative prose are located in the positive range along Axis 1. And as is shown in Table 2, this axis explains more than 65% of the information contained in the frequency table. So it may be safely said that the major factor which determines the use of modals across genres can be attributed to the imaginative vs. informative dichotomy. This dichotomy has always been the principal one whether from the viewpoint of the distribution of pronouns as attested in Nakamura (1989b) or from the viewpoint of the distribution of grammatical tags as attested in Nakamura (1990). The contrast between the two major categories is thus objectively testified in case of the modals, as well.

In imaginative prose, Genre R (Humour) is located rather close to the origin, separated from the rest, and this was also the case in Nakamura (1989b) and Nakamura (1990). Humour in the Brown Corpus thus assumes a unique position in imaginative prose. The five genres of imaginative prose do not show high quantities either along Axis 2 or along Axis 3, which is also the case in Nakamura (1990). In informative prose, the genres are all scattered around in the negative range of Axis 1 except for Genre A, which assumes a positive value along this axis. Genres E (Skills and Hobbies), D (Religion), and H (Government documents) are located at the extremities of Axis 1, and Genres E and H are sharply in contrast to each other along Axis 2. The rest of the genres are rather close to one another and located around the center of the group.

The reason why the genres are distributed as in Figure 1 can be explained by referring to the distribution of modals shown in Figure 2. Those modals which are abundant in imaginative prose and scantily used in informative prose are located in the positive range along Axis 1, assuming more or less the same position as the genres in imaginative prose in Figure 1.[5] The typical example is *would*. Thus, the locations of modals in Figure 2 correspond to the positions of genres in Figure 1, and *vice versa*. In other words, Figure 2 can be used to explain the genre distribution in Figure 1, or Figure 1 can be used to explain the modal distribution in Figure 2.

A group of the modals *would, used, ought,* and *might* can be said to be rather imaginative-prose-oriented, while another group of modals containing *dare, need, may, shall,* and *could* can be said to be informative-prose-oriented. Yet another group of modals containing *can, should* and *must* is located rather close to the origin but still in the negative domain of Axis 1, indicating that they are oriented a little towards informative prose but not very much. Typically neutral along Axis 1 is the modal *will.*

Along Axis 2, the high frequency of *shall,* which is found in Genre H, and that of *could,* which is found in Genre E, are in contrast. Genre D, which is religion, assumes the closest location to Genre H because of the abundance of *shall* in the Biblical quotations. Likewise, Genre A is characterized by the high frequency of *will.* In this way, the structure of the Brown Corpus from the viewpoint of the use of modals has been determined as in Figure 1 and Figure 2, revealing the ways modal auxiliaries are used in various genres of the corpus.

3.2. The LOB Corpus

Figure 3 indicates that, in case of the LOB Corpus, genres are also neatly separated between informative and imaginative prose except for one case, i.e. Genre G (*Belles lettres,* biography, essays), which is located in the negative region along Axis 1 where the genres of imaginative prose are located. Genre G, in fact, is the closest to the origin of coordinates, indicating that it does not show much preference for or against the use of modals.

Among the genres of imaginative prose, Genre R (Humour) and Genre M (Science fiction) assume locations rather closer to the origin of coordinates. In other words, they are rather close to informative prose. This tendency is also shown in Nakamura (1991b), which examined the distribution of the grammatical tags. Science fiction in the LOB Corpus is also separated from the rest of the genres of imaginative prose. As can be seen from Figure 4, the modals which characterize imaginative prose turn out to be *dare, could, used, might, ought* and *would.*

The genres of informative prose are scattered around in the positive range of Axis 1, but it seems that the modals *may, should, shall,* and *will* are the ones that characterize informative prose. *Can* and *need* are also rather informative-prose-oriented. Axis 2 contrasts Genre H on the one hand and Genres C (Press: reviews), E, A and possibly B (Press: editorial) on the other. This contrast matches the one between *shall* and *will* as seen in Figure 4.

Another thing to be noticed in Figure 4 is that *must* is quite neutral especially along Axis 1, indicating that many genres in the LOB Corpus do not show much preference for or against the use of this modal. Thus, the structure of the LOB Corpus based upon the distribution of modals has also been determined as in Figure 3 and Figure 4.

3.3. Comparison between the Brown and the LOB corpora

Several similarities and differences are to be noted between the Brown and the LOB corpora from the foregoing observations. For one thing, the genres of informative and imaginative prose are clearly separated in both of the corpora as seen in Figure 1 and Figure 3. The genres H and E of informative prose are in contrast to the genres in imaginative prose along Axis 1 and are in contrast to each other along Axis 2. But these seemingly similar distributions are somewhat coincidental since they are not based upon exactly the same kind of distributions but rather on different distributions of modals as can be seen from Figures 2 and 4. For example, the contrast between H and E in the Brown Corpus is attributed to the contrast between *shall* and *could*, while the same contrast in the LOB Corpus is ascribed rather to the one between *shall* and *will*.

The greatest difference can be found in the uses of *dare* and *could*. These two modals behave in completely opposite ways along Axis 1. In the Brown Corpus, they are the ones characterizing informative prose while, in the LOB Corpus, they are used more abundantly in imaginative prose than in informative prose. Perhaps the most important difference lies in the behaviour of *could* since the number of occurrences of *dare* is very small. Another large difference is the position of *would*. In the Brown Corpus, *would* is the main modal that characterizes imaginative prose, but in the LOB Corpus it is located nearest to the origin among those characterizing imaginative prose; its use is not really typical of imaginative prose in the LOB Corpus.

Another major difference is found in the use of *will*. In the Brown Corpus, *will* occupies a rather neutral position along Axis 1; it is neither characteristic of informative nor of imaginative prose in general.[6] But in the LOB Corpus, it is located toward informative prose. Genre E, as mentioned in passing above, and Genre A are characterized by an abundant use of this modal.

3.4. The combined corpus

It turned out to be the case that there is a good deal of difference

between the Brown and the LOB corpora and the difference can best be seen if these two corpora are combined, and if quantification is conducted again on this combined corpus. So the figures in Table 1 were adjusted again to the average genre size for the two corpora, and fed into Hayashi's Quantification Method Type III once more. As can be seen from Table 5, the cumulative proportions accounted for which were obtained this time were lower than the ones obtained for either Corpus individually, but this seems to be inevitable. Naturally, the larger the number of samples gets, the less clear the overall hidden tendencies become. But a cumulative proportion close to 80% can still be evaluated as being very high. The quantities given to genres and modals are shown in Table 6 and Table 7 and plotted in Figure 5 and Figure 6. (In Figure 5, the corpus is identified either by B (the Brown Corpus) or L (the LOB Corpus) placed before alphabetical letters indicating genres.)

Table 5. Correlationship Coefficients, Proportion Accounted for, and Cumulative Proportion Accounted For for Each Axis.

Corpus	Axis	Correlationship Coefficients	Proportion Accounted for	Cumulative Proportion Accounted For
Combined	1	0.3074206D+00	0.4412826D+02	0.4412826D+02
	2	0.2130426D+00	0.2119259D+02	0.6532085D+02
	3	0.1748106D+00	0.1426878D+02	0.7958963D+02

As shown in Figure 5, the major division between informative prose and imaginative prose can still be seen along Axis 1, which explains about 44% of the information contained in the adjusted frequency table for the combined corpus. The only exceptions to this are BA and LA, located in the positive region along this axis. Genre A both in the Brown and the LOB corpora is closer to imaginative prose than the rest of the genres in informative prose. Major modals contributing to this dichotomy are *may* and *would*.

Two items that showed quite opposite distributions along Axis 1 when the two corpora were separately analyzed above, i.e. *dare* and *could*, now appear to be the major factor of Axis 2 as seen in Figure 6. And this fact is reflected in the separation of imaginative prose in the Brown and the LOB corpora along this axis in Figure 5. Another thing to be noted here is that *could* is rather informative-prose-oriented on the whole along Axis 1, making all the genres in imaginative prose in the LOB Corpus much closer to informative prose than those in the Brown Corpus.

Table 6. Quantities Given to Genres.

Corpus	Genre	Axis 1	Axis 2	Axis 3
	A	0.1635139D+00	0.8976886D+00	0.2176875D+01
	B	-.2765414D+00	-.1216082D+00	0.9370405D+00
	C	-.5379832D+00	-.4835516D+00	0.4454373D+00
	D	-.1160131D+01	-.8215395D+00	-.7155088D+00
	E	-.1132395D+01	-.1187781D+01	0.1937605D+01
	F	-.6239452D+00	-.6247864D+00	0.1766823D+00
	G	-.1964125D+00	-.6197080D+00	-.3421663D+00
(BROWN)	H	-.1452367D+01	0.7353463D+00	-.1289461D+01
	J	-.8400826D+00	0.2945420D+00	-.1652999D+00
	K	0.2136018D+01	0.9862183D+00	-.5625663D+00
	L	0.1806050D+01	0.8368563D+00	-.5850241D-01
	M	0.1848141D+01	0.9315848D+00	-.3233562D+00
	N	0.1661665D+01	0.1053359D+01	0.2662620D+00
	P	0.1547930D+01	0.7474589D+00	0.2936163D+00
	R	0.1131840D+01	0.3635322D+00	0.8523403D-01
Combined				
	A	0.5225226D-01	0.8408085D+00	0.1034579D+01
	B	-.4204260D+00	0.7287610D+00	0.7380811D+00
	C	-.5702128D+00	-.4301578D+00	0.1335840D+01
	D	-.5938456D+00	0.8053661D+00	-.5742663D+00
	E	-.8292224D+00	0.1063799D+01	0.1287859D+01
	F	-.6137426D+00	0.2949087D+00	0.4705909D+00
	G	-.9207755D-01	-.3819908D+00	-.5473762D+00
(LOB)	H	-.1224491D+01	0.1548449D+01	-.2500933D+01
	J	-.1065013D+01	0.7478438D+00	-.7637576D+00
	K	0.8207535D+00	-.1260846D+01	-.9410511D+00
	L	0.7726316D+00	-.2366222D+01	-.6783103D+00
	M	0.1309873D+00	-.8680251D+00	-.5225405D+00
	N	0.4446337D+00	-.2179674D+01	-.2113085D+00
	P	0.5849079D+00	-.1128596D+01	-.8046322D+00
	R	0.5061963D+00	-.3653545D+00	0.3374936D+00

Table 7. Quantities Given to Modals.

Corpus	Modal	Axis 1	Axis 2	Axis 3
	CAN	-.4926516D+00	-.2564827D-01	0.6497206D+00
	COULD	-.2854124D+00	-.2359285D+01	0.2446620D+00
	MAY	-.1780285D+01	0.6109857D+00	-.7807393D+00
	MIGHT	0.9347327D+00	-.4373267D+00	-.7449034D+00
	MUST	-.3358366D+00	-.2616132D+00	-.6678998D+00
	OUGHT	0.1012782D+01	-.5660212D+00	-.1016493D+01
Combined	SHALL	-.1451496D+01	0.1020596D+01	-.4302163D+01
	SHOULD	-.8152997D+00	0.6710289D+00	-.2322402D+00
	WILL	-.1738506D+00	0.8658490D+00	0.1377957D+01
	WOULD	0.1523645D+01	0.3637431D+00	-.4199853D+00
	NEED	-.9145795D+00	-.4145599D+00	-.6804618D+00
	DARE	0.1243318D+00	-.5374785D+01	-.2725513D+01
	USED	0.1086319D+01	-.9559976D+00	-.5613694D-01

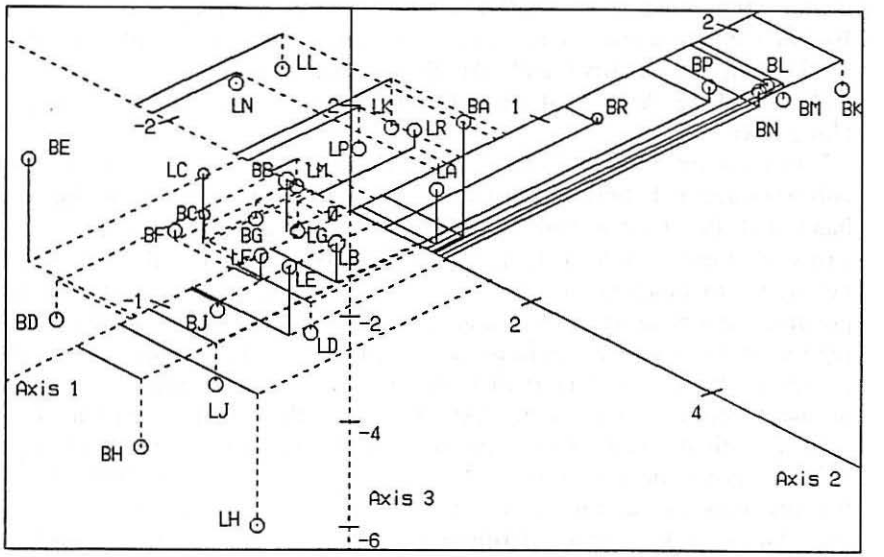


Fig. 5. Three-Dimensional Distribution of Genres in the Combined Corpus.

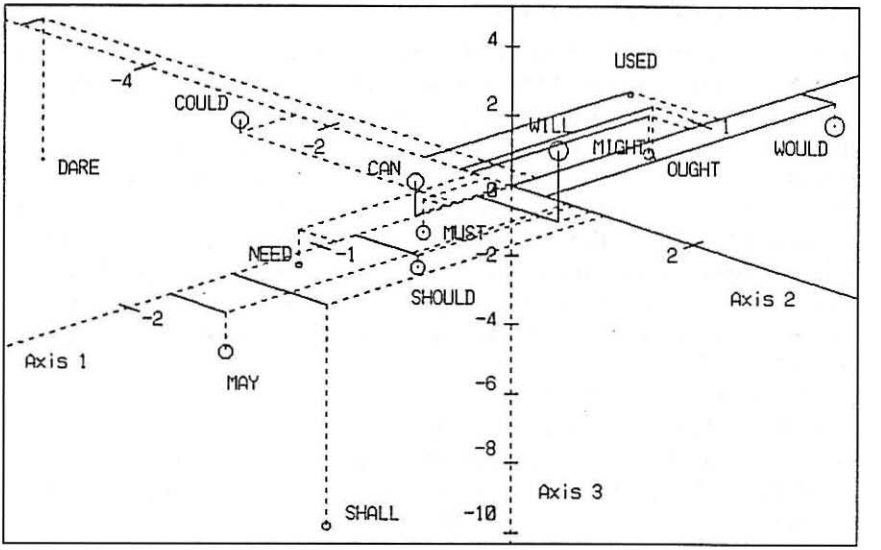


Fig. 6. Three-Dimensional Distribution of Modals in the Combined Corpus.

The contrast between *shall* and *could* observed along Axis 2 of the Brown Corpus seems to be subsumed along Axis 2, too, placing *shall* at the extreme positive end. Another contrast observed between *shall* and *will* along Axis 2 of the LOB Corpus now appears to be shown along Axis 3.

When Figure 5 is studied carefully, there appears to be an interesting correspondence between informative and imaginative prose on the one hand and the Brown and the LOB corpora on the other barring the cases of Genres BA, LA, LC, BH and BJ. That is, all the genres belonging to imaginative prose in the Brown Corpus are located in the positive region along Axis 1 and Axis 2 with the rest of genres in the region of the opposite polarity along these two axes, whereas all the genres pertaining to imaginative prose in the LOB Corpus are located in the positive region along Axis 1 and in the negative region along Axis 2 with the rest of the genres again in the region of the opposite polarity along these two axes. As might be inferred from Figure 6 and the previous discussion, it is the use of *could*, *would* and *will* that contributes to this neat distribution of genres. Therefore, it could be safely concluded that these three modals play a vital role in distinguishing the Brown and the LOB corpora.

4. Concluding remarks

The occurrences of thirteen modals were counted across genres for the Brown Corpus and the LOB Corpus and were fed into the statistical procedure called Hayashi's Quantification Method Type III. As a result, the structures of the two corpora from the viewpoint of the use of these modals have been determined as shown in Figure 1 through Figure 4. In so doing, the relationships among fifteen genres in each corpus, the relationships among thirteen modals and more importantly the relationships between genres and modals, have been made explicit quantitatively. The major factor which plays an important role in determining the structures turns out to be the informative vs. imaginative dichotomy, as was expected.

It might be expected that the use of important grammatical categories like the modals would not be as different in respect of either American or of British English since grammatical categories seen to be quite resistant to change in contrast to vocabulary items (which may be more susceptible to changes through time as demonstrated in Hofland and Johansson (1982)). But the truth is that the use of some of the modals does differ markedly between American and British English as revealed

by our comparison of the two corpora. The fact was made more explicit especially when the combined corpus was processed again by the same statistical procedure. A study of Figure 5 and Figure 6 led to the conclusion that the major difference between the two corpora can be attributed to the usage of *could*, *would*, and *will*.

Now that the present study is at its end, one can see the usefulness of the method employed here. But there definitely is a limit to this kind of quantitative methodology. For one thing, the axes of the figures are not only used for grouping and separating the items as in the present study, but they could also be given proper interpretations in many cases. In the present study, Axis 1 was interpreted as reflecting the informative vs. imaginative dichotomy, but it seems very difficult to give proper interpretations to Axis 2 and Axis 3. What is more, although the fact that certain modals are closely related to certain genres could be made clear, how these modals are actually related to the genres is not made explicit at all. This is a limit of a quantitative study like the present one; qualitative studies, investigating the actual occurrences of the modals in the texts, should follow up the present study.

Notes

1. This work was supported by a grant from the Japanese Ministry of Education for studying abroad for the fiscal year of 1991. The author hereby wishes to express his gratitude to the Norwegian Computing Center for the Humanities, which provided him with the corpora together with the software and hardware to handle them. He also wishes to express his gratitude to Knut Hofland of the above center for his patient assistance, without which this work would never have been completed.
2. Genre sizes and corpus sizes are found in Francis and Kučera (1982:533) and Johansson and Hofland (1989:7).
3. Those who are interested in this technique and want to know about it in more detail are asked to refer to a standard textbook on statistics, but as this technique was developed by a Japanese statistician, it may be difficult to find it in the textbooks written in English or other European languages. Most of the Japanese textbooks, including the one in the references, take it up. It is also available as a statistical package both for main-frame computers and personal computers in Japan.
4. Polarities of Axis 1 and Axis 3 are reversed in this figure so that

the comparison between the two corpora can be effectively conducted. As the quantities given to genres and modals determine only the relative positions, the reversal of the polarities of axes does not affect the validity of the interpretations of figures.

5. The abundance and scarcity of modals are not shown directly in the original or adjusted frequency tables but they are obtained by examining the discrepancy found between the actual observed frequency and the hypothetical frequency expected if the modals were distributed randomly, or evenly, irrespective of genres. This hypothetical frequency is the same as the expected frequency used in the chi-square statistics. One can draw figures showing these discrepancies genre by genre or modal by modal, using a portion such as 10% or 20% of the expected frequency as a unit as found in Nakamura (1989b) and Nakamura (1990), but they are omitted here.
6. Although Genre A of informative prose is highly characterized by the use of this modal, this fact appears not on the first axis but on the third axis.

References

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

Biber, Douglas. 1989. A typology of English texts. *Linguistics* 27:3-43.

Francis, W. Nelson and Henry Kučera. 1982. *Frequency analysis of English usage: Lexicon and grammar*. Boston: Houghton & Mifflin.

Hofland, Knut and Stig Johansson. 1982. *Word frequencies in British and American English*. Bergen: Norwegian Computing Center for the Humanities.

Johansson, Stig. 1979. Corpus-based studies of British and American English. In *Papers from the Scandinavian Symposium on Syntactic Variation*, ed. by Sven Jacobson. 85-100. Stockholm: Almqvist & Wiksell.

Johansson, Stig and Knut Hofland. 1989. *Frequency analysis of English vocabulary and grammar based on the LOB Corpus*. Vol. 1: *Tag frequencies and word frequencies*. Vol. 2: *Tag combinations and word combinations*. Oxford: Clarendon Press.

Kučera, Henry and W. Nelson Francis. 1967. *Computational analysis of present-day American English.* Providence: Brown University Press.

Mizutani, Sizuo. 1997. *Suri-Gengogaku* (Mathematical Linguistics). Tokyo: Baifu-kan.

Nakamura, Junsaku. 1985. On the methodologies of quantitative groupings of English texts. *JACET* (The Japan Association of College English Teachers) *Bulletin* 16:133-148.

Nakamura, Junsaku. 1986. Classification of English texts by means of Hayashi's Quantification Method Type III. *Journal of Cultural and Social Science, College of General Education, University of Tokushima* 21:71-86.

Nakamura, Junsaku. 1987. Notes on the use of Hayashi's Quantification Method Type III for classifying English texts. *Journal of Cultural and Social Science, College of General Education, University of Tokushima* 22:127-145.

Nakamura, Junsaku. 1989a. Creation of a vocabulary frequency table from the Brown Corpus. *Journal of Cultural and Social Science, College of General Education, University of Tokushima* 24:171-182.

Nakamura, Junsaku. 1989b. A quantitative study on the use of personal pronouns in the Brown Corpus. *JACET Bulletin* 20:51-71.

Nakamura, Junsaku. 1990. A study on the structure of the Brown Corpus based upon the distribution of grammatical tags. *Journal of Foreign Languages and Literature, College of General Education, University of Tokushima* 1:13-35.

Nakamura, Junsaku. 1991a. A study on the structure of the Brown Corpus based upon the distribution of its vocabulary items. *Journal of Foreign Languages and Literature, College of General Education, University of Tokushima* 2:27-47.

Nakamura, Junsaku. 1991b. The relationships among genres in the LOB corpus based upon the distribution of grammatical tags. *JACET Bulletin* 22:55-74.

Nakamura, Junsaku. 1992. The comparison of the Brown and the LOB corpora based upon the distribution of grammatical tags. *Journal of Foreign Languages and Literature, College of General Education, University of Tokushima* 3:43-58.

Nakamura, Junsaku. In preparation. Hayashi's Quantification Method Type III: A tool for determining text typology in large corpora.

Nakamura, Junsaku. In preparation. Statistical methods and large corpora: A new tool for describing text types.

Sinclair, John M. 1991. *Corpus, concordance, collocation*. Oxford: Oxford University Press.

Yasuda, S. and M. Unno. 1977. *Shakai Tokeigaku* (Social statistics). 2nd rev. ed. Tokyo: Maruzen.

Adjectives and nouns with reported clauses[1]

Jacques Noël
University of Liège

1. Introduction

In this progress report I return to the problem of adjective and noun complementation by a *that*-clause, or, to use the terminology of the COBUILD grammar and dictionary (Sinclair *et al.* 1987, 1990) adjectives and nouns with reported clauses. Using checklists of adjectives and nouns established in previous, unpublished studies, I will try to extend these checklists by finding attested examples of adjectives and nouns whose use with reported clauses has so far gone unnoticed or unrecorded in our reference grammars and dictionaries. The two checklists are appended: Appendix 1 (Adjectives), Appendix 2 (Nouns).

Shortcomings in the treatment of the subcategories under discussion are apparent in our learners' dictionaries, no matter how good these are on verbs. Specifically, I am thinking of OALDE 1974 and 1989, LDOCE 1978 and 1987, and COBUILD 1987 and 1989, a Collins-Klett edition; of these, the earlier edition in each case is available to our Liège team in machine-readable form, but only the online COBUILD is used systematically in this paper. Reference grammars, as we know, also tend to regard adjective and noun complementation by a *that*-clause as somehow marginal. In this borderline area between lexicography and grammar, there is in fact precious little to go by, for a foreign learner and for natural language processing applications.

Learners' dictionaries, when they do address the problem, are very sketchy and unsystematic. As you may remember, adjective and noun patterns are given some attention by Hornby in his grammar, but not in his dictionary, and this is still the case in the new OALDE. The new LDOCE is better on adjective and noun complementation than the

first edition. It has systematically replaced code Number 5 by a *+that* feature attached to all parts of speech, and it assigns the appropriate code to each example, but many adjective and noun examples remain unrecorded (*acceptance, admirable, alarmed, alarming*) and/or uncoded (*amazed*). As for COBUILD, the first edition assigns the code REPORT-CL to less than fifty adjectives and roughly as many nouns, whereas each of my checklists (based mainly on various dictionaries and corpora) contains over two hundred items. While it has many new and interesting features (statistics on word usage in definitions and on grammar codes), the Collins-Klett COBUILD has given up the subcategory altogether for adjectives and nouns.

As far as reference grammars are concerned, let me limit myself to Quirk *et al.* 1985 (and Greenbaum & Quirk 1990, referred to later) and to Sinclair *et al.* 1990, the COBUILD grammar.

Quirk *et al.* deal with adjective and noun complementation in a chapter on verb complementation. In a section on 'Complementation of abstract nouns', with references to other sections on nominalization and apposition, they focus on nouns morphologically related to verbs and adjectives, insisting on the fact that correspondence in complementation is not automatic. In the sections on adjective complementation by a *that*-clause, the main distinction is between adjectives with experiencer, and those with anticipatory *it* as subject, and further subcategorizations are based on mood (indicative, subjunctive, putative *should* in the *that*-clause), and on the nonparticipial vs participial distinction. I have tried to use the generalization about participles in order to extend my checklist of adjectives compiled from dictionaries and corpora. As to the grammatical contexts (such as '*it is* + adjective + *that*') favoured by the adjectives and nouns I am interested in, I will also refer to Greenbaum and Quirk 1990.

Another area in which the 'Comprehensive' and the 'Student's Grammar' throw light on our problem has to do with theme, focus, and the contribution of extraposed subject clauses and others, by end weight and focus, to the communicative function of 'bringing ... an entire proposition to the attention of the hearer' (Greenbaum and Quirk 1990: 424). Though I have no solid evidence of this yet, I have an impression that adjectives and nouns with reported clauses are typical of the serious newspapers (whether British or American); typically, the popular press plays on short sentences which, by their very nature, provide more end focuses, and it therefore tends to prefer alternative constructions, as in:

(1) But one snag could delay any decision: rivalry among Yeltsin insiders for the top espionage job. (*Newsweek*)

Lastly, let us bear in mind that some disjuncts in the Greenbaum sense are basically report structures of the type examined here: 'The thing is...', 'The good news is that... The bad news is that...', 'One is that... The other one is that...'.

The COBUILD grammar takes a fresh look at our problem. Under the broad heading of 'report structures', Chapter Seven ('Reporting what people say or think') offers a unified treatment of those verbs, nouns and adjectives that introduce a 'fact', piece of knowledge, etc., 'reported' or described in a following *that*-clause. The grammar elaborates on the notion of REPORTED clause (cf the dictionary's REPORT-CL code) typically used with nouns referring 'to what someone says or thinks', or to 'facts', 'beliefs', or 'ideas', and with adjectives with *that*-clauses describing a feeling ('X is sad that...') or 'the cause of the feeling' ('it is sad that...'), or 'indicating knowledge' ('aware'), or a 'fact' commented on ('true'). Other relevant passages are:

- The final points in Chapter 7 on nominal use of *that*-clauses: 'the fact that... is a trivial irrelevance' is described as the equivalent of 'it is irrelevant that' in less formal English.
- A discussion in Chapter 10 on ways of 'Commenting on an action, activity, or experience' (no *that*-clause) vs 'Commenting on a fact that you are about to mention': adjectives (eg *amazing*) and nouns (eg *disgrace*) typically with *that*-clause (*ibid*, 413-415).
- Prefacing structures 'which point forward to what [people] are going to say and classify or label ... in some way', and which may reach beyond the sentence into discourse as in 'It was rather funny:...' (*ibid*, 429-430).

Examples like the following (2) show the relationship between reported speech and the *that*-clauses we are concerned with here, thus justifying the COBUILD terminology:

(2) Report structures

(a) The rule of thumb in Toronto is that if you are black and you want to be harassed by the police, drive an expensive car (*The Economist*)

(b) An old Scottish saw is that when you visit someone in Glasgow, you will be greeted: "Come away in, you'll be wanting your tea!"

In Edinburgh, you will hear: "Do come in! You'll have had your tea?" (*The Economist*).

The property under discussion is not safely predictable from derivation: contrast 'the accusation *that...*', and 'he accused her *of...*' as well as 'he stands accused *of...*'. I have not attempted to pursue this matter here.

Furthermore, there are adjectives (*adamant*) and nouns (*idea*) that have no cognates with the same syntactic behaviour. Nor are our report structures safely predicted on a semantic or thesauric basis. But despite these difficulties I have attempted to use keywords extracted from COBUILD's definitions and thesaurus to identify unrecorded report structures in corpora. This is discussed in the next two sections: 2. Using keywords from definitions and 3. Using keywords from a thesaurus. As the results of both of these strategies (even in terms of the number of 'new' items uncovered) are not all that good, I developed a number of computational strategies, which turn out to be much more productive. Based mainly on grammatical contexts (such as 'it is ... that') favoured by the subcategories under discussion, the filtering, sorting, concordancing, and searching operations used and their pros and cons are discussed in the fourth and last section: 4. Using grammatical patterns to search new corpora.

My intention had always been to use much larger corpora than those I had used so far, and I am pleased to report that I have been able to do so on a scale that I had not thought possible only a year ago. For one thing, the emergence of CD-ROM corpora largely took place in the past twelve months or so, and, in addition, our Department was able to purchase the academic licence for use of the New OED file. From this very large SGML file (more precisely an ASCII file of close to 550 MB, prepared by Jacques Jansen from the original tapes), I extracted a file of all the twentieth-century examples. In addition, I was able to use a whole year of *The Independent* on CD-ROM (1989-90) and three years of the *Wall Street Journal* (1987-89) from the ACL/DCI CD-ROM (September 1991). From these three corpora, I extracted all the passages in which the word *that* occurs. Extracting, and then processing such a large mass of citations (over 100 MB of text) on a PC had also become possible; last year, our Department had purchased a rewritable optical disk drive with cartridges which, thanks to the STACKER compression software (also a newcomer on the 1991 market), can store a maximum of 250 MB. It was of course essential to avoid spending ages browsing through such large corpora visually line by line,

hence the various strategies, lexicographic, grammatical, and computational, reported in this paper.

2. *Using keywords from definitions*

For a number of reasons, the approach based on using keywords from definitions does not work, as it retrieves either too much (it produces noise) or too little (it results in silence). After some dips into the files of LDOCE 1978 and OALDE 1974, I had to conclude that their definition language was not likely to provide usable keywords, so I decided to concentrate on COBUILD, encouraged by the fact that it represents a unique attempt to rethink and systematize definition-writing.

Let me now explain why even keywords from the COBUILD definitions did not fulfil expectations. First of all, there is of course one obvious reason for 'silence', and this is when the relevant word is not entered as such in the dictionary.

(3) Example: **grouch**
Part of his grouch was that he had tried to join the Air Force and always been put off (New OED). COBUILD only enters the noun *grouch* meaning 'someone who is always complaining...'

But the main reason is that the attempted formalization or normalization does not go far enough; the COBUILD definitions are still too opaque, or still lack transparency in at least three crucial respects:

(i) LEXICAL OPACITY. The definition does not contain any keyword likely to signal or predict the reported-clause feature (consider *something* in the following definition), or at least to do so unambiguously, because the syntactic environment (as explained below) of a possible keyword is unrevealing, uninformative, irrelevant, or, worse still, misleading, with respect to predicting the feature we are interested in; also consider 'to lose interest or enthusiasm', all words which rule out a reported-clause structure.

(4) Example: **turn-off**
But the biggest turn-off is that... (from a text on lack of long-term research investments by the private sector, *The Economist*). COBUILD definition: Something that is a turn-off causes you to lose interest or enthusiasm.

(ii) SEMANTIC OPACITY. Despite the claims and merits of COBUILD, its definitions cannot, even in part, be used to retrieve even those very words that are duly coded REPORT-CL in the dictionary, not to mention

a larger subset of the vocabulary having the property in question. Though alphabetical sorts of the definition texts display a number of fascinating regularities, the formalization does not go far enough: Appendix 3 gives alphabetically sorted definitions which do show a limited amount of regularity in definitions of words with REPORT-CL (the opening ten lines of the relevant adjective and noun definitions). But I think I am correct in concluding that the definitions would all have to be rewritten in a normalized fashion if they are to be used successfully in the kind of filtering or retrieval operations that I am interested in here. Needless to say, a definition parser would need to have access to the very information that this paper attempts to uncover.

Using one-word keywords extracted from the definitions (words like *promise*, *certain*, etc) often produces far too much noise to be useful. This is due to the syntactic complexity of the COBUILD definitions, and sometimes, to the frequency of the keywords in question (see figures appearing in Cobuild-Klett).

In all fairness to the COBUILD editors, however, let me give one example of a very innovative COBUILD definition which, at least for a human user, does offer some semantic transparency, in that it comes close to expressing the causal link that the grammarian identifies in the generalization expressed in the COBUILD grammar. In John Sinclair's words, discussing a type of adjective like *afraid*, *sad*, etc: 'if you want to say what causes someone to have a particular feeling, you can mention the cause of the feeling in a *that*-clause after an adjective describing the feeling'. Contrast the semantics of '(s)he is sad' and of 'it is sad that', where the adjective 'comment(s) on a fact' reported in the *that*-clause. The definition of *satisfied* in the dictionary reads: 'Someone who is **satisfied** is 1 happy because they have got what they wanted' (not coded for REPORT-CL, though the use of *happy* as synonymous genus word does not rule out this structure); '2 convinced that something is true or settled' (the REPORT-CL is duly coded and exemplified here, even if contrary to Definition 1: '...happy because...'). Definition 2 unfortunately fails to make explicit the causal link between the 'feeling' referred to and the state of affairs described in the *that*-clause.

(iii) SYNTACTIC OPACITY. For computers and learners alike, a first, most obvious, difficulty is the ambiguity of the word *that*, sometimes used as conjunction and as relative pronoun in the same set of definitions, occasionally even in very difficult structures with 'pushdown *wh*-element'

as subject. Compare Quirk *et al* (1985: 821f, 1298) and the first definition below:

(5) (a) I will read the memo (which) [that:JN] Pat hopes (that) John will send you

(b) I will read the memo (which) [that:JN] Pat hopes (*that) will be sent to you

(6) COBUILD definition:

1 A conclusion is something that you decide is true as a result of knowing that other things are true.

(7) COBUILD definitions:

2 Something that is a guarantee of something else makes it certain that it will happen or that it is true.

5 A guarantee is also 5.1 a promise that you will do something, or that something will definitely happen.(...).

5.4 a written promise by a company that if a product that they sell or work that they do has any faults within a particular time, it will be repaired, replaced, or redone free of charge.

The main point about syntactic transparency of definitions is that, ideally, at least one clearly identified defining word (definiens) should have the same syntactic property or properties as the word(s) defined (or definiendum). For example: Webster's Third uses the syntactically cognate adjective *insistent* in its definition of *adamant* (only LDOCE 1987 aptly labels examples of both of these adjectives *+that*). Similarly, OALDE 1974 defines 'have a **hunch** that' by 'think it likely that'.

Examples of COBUILD definitions that neatly capture contrasts in complementation are the following:

(8) If someone is anxious to do something or anxious that something should happen, they very much want to do it or very much want it to happen.

(9) If you are aware that something such as an important problem or difficulty exists or if you are aware of it, you know about it, either because you have thought about it or because you have just noticed it.

(10) If you are aware of something or aware that something exists or is happening, you realize it because you hear it, see it, smell it, or feel it.

On the other hand, consider this definition of *ban* which wrongly

predicts, and suggests to a inexperienced learner, that the word in question can have complementation by *that*-clause:

(11) A ban is an official statement that something must not be done, shown, used, etc.

3. *Using keywords from a thesaurus*

To illustrate the use I made of the COBUILD thesaurus in this study, let me take the beginning of the list of terms I used for adjectives. These terms appear in the columns of the printed dictionary. Words with one asterisk (eg **afraid*) are those already on my checklist, those with two asterisks (eg **controversial**) are those that are not, and are considered likely candidates for the REPORTED-CL feature, and which were therefore included in a new checklist (let us call it 2A). Needless to say, particularly for a non-native speaker like myself, the judgements involved are sometimes difficult to make, and mistakes in such a partly manual (i.e. not fully automatic) procedure are likely. For instance, I failed to assign keyword status to *inadvisable*.

(12) Entries from the COBUILD thesaurus (Liège version):
accepted ANT *controversial*
adamant SYN resolute ANT hesitant, pliable
admirable BT *good SYN *excellent* SYN splendidly
advisable SYN *wise* ANT *foolish* , *inadvisable ANT folly
afraid SYN *frightened, *scared*
afraid SYN *scared*
afraid BT unfortunately
alarmed BT *afraid
alarming BT *worrying SYN *disquieting*

As is to be expected, the method reflects the inadequacies of the thesaurus used; let me stress that the COBUILD thesaurus is not any better or worse for our purposes than others. Every time there is a gap in the COBUILD thesaurus, you can of course only expect the search to produce silence. One such gap is *boisterous* (ANT: docile; defined as 'noisy, lively and full of energy' in COBUILD), of which the following example appears in OED2, in the entry for *cherry* (in the sense of 'a virgin'):

(13) **boisterous**
She now held off my hands and now led them inside her dress, alleging instruction, boisterous that I was still cherry (New OED)

The list of words of Checklist 2A which, on the basis of thesauric information gathered as I have just explained, were found to have an attested example with reported clause is appended (Appendix 6).

Because of the low productivity of this strategy, I decided not to test its productivity for nouns, not to mention multi-word items (also discussed below), even though I have appended a noun list obtained from the COBUILD thesaurus. This low productivity is due not only to gaps in the thesaurus used, but perhaps even more to the fact that semantic and thesauric relations cannot reliably predict syntactic behaviour (whereas the converse is not true, as I now wish to argue). There is also the fact that, like in other such works, no attempt was made in the COBUILD thesaurus to associate grammatical information with its terms.

4. *Using grammatical patterns to search new corpora*

The use, in my previous work, of corpora of dictionary examples (from COBUILD, among others) and of classic text corpora (in particular the tagged version of the LOB Corpus) produced fairly good results, in terms of increase in the number of items on my checklists (from some fifty items from COBUILD to over two hundred in each list). But on various occasions examples of 'new' adjectives and nouns with reported clauses kept cropping up in my readings, often not entered as such, or even with no relevant citation, in COBUILD, and in other dictionaries. Even worse, as neither definition nor synonyms seemed to be reliable predictors of report structures, there seemed to be no principled basis on which to search for the property in question. Here are two citations, discovered by pure chance.

(14) **emphatic**
Customs and Excise was yesterday emphatic that ...
(*The Independent*)
COBUILD Thesaurus: SYN insistent

Unfortunately, *insistent* is not assigned the REPORT-CL label either, and the property cannot be predicted from its superordinate *resolute*.

(15) **jibe/gibe**
The common jibe against Palestinians is that ...
(*The Economist*)
COBUILD Thesaurus: jibe BT remark SYN insult
COBUILD definition: A jibe is a rude or insulting remark...

It soon became apparent that a very large corpus was called for, hence my decision, mentioned earlier, to produce what turned out to be over one hundred megabytes of citations with the word *that*, extracted from the New OED examples, *The Independent*, and the *Wall Street Journal*. In fact, it turned out, as I wish to argue, that the most useful basic resource for my project was a range of concordances with the word *that* as primary sort key. What proved to be particularly useful was the Kwic instruction in the CLAN/CHILDES package, with -c111 characters of context; only rarely does this 111-character 'window' leave out a head adjective or noun I am interested in, but this was not sufficient reason to give up this convenient format; otherwise I would have had to operate with still larger windows of one line before and one line after (as provided for by a Kwal-instruction in CLAN/CHILDES). The MKS and the Thompson Unix/Awk utilities were of course also used extensively. Table 1 gives an example, listing the first ten records, in alphabetical order, in the *that*-concordance produced for *The Independent*; my program has added a copy of the word preceding *that* as Field One of the record, that is, 'fronted' the word in question (things went wrong with quotes, etc, in the example given here).

Table 1: KWIC concordance, sorted alphabetically after 'fronting' of the 'word' preceding 'Ithat'

, ° 91171| endorsement of threats to Salman Rushdie's life
, Ithat ours is indeed, as is often somewhat glibly said,
A ° 380789| untruth. He added: 'I don't think with
Mrs A Ithat it was fantasy, that was more deliberate untruth
A ° 433179| delegates and were bona fide.//// A 'TACIT
ADMISSION' Ithat Special Air Service officers have been training Cambodian
AA ° 107753| day's weather had// road. Hampshire police told
the AA Ithat almost every major route in the county was affected
AAC ° 14449| near term'. UAL said that it had been advised by
AAC Ithat syndication of the financing would be possible on
AAIB, ° 237481| although not necessarily the view of
the AAIB, Ithat three pilots are safer than two,' he said. Pilots
A ° 338635| social comedies (Abigail's Party, Smelling A
Rat) Ithat sketch the class-riddled, emotionally inarticulate
ADAGE ° 198531| and investigative expertise to////
THE ADAGE Ithat attack is the best form of defence was adopted
yesterday
AGM ° 8675| manager, yesterday promised the shareholders'
AGM Ithat the team will not be relegated and will climb from

Second, and most important, in order to facilitate browsing of this

prohibitively large corpus, I decided to make use, as secondary, tertiary, etc, search and/or sort keys, of grammatically or syntactically defined strings, which favour the use of nouns and adjectives with reported clauses. Contexts favouring N+*that* or Adj+*that* are:

- existential *there* sentences
- negative contexts: '(there is/was/etc) no ... that...'
- cleft: 'What is ... is that...'
- pseudo-cleft: 'it is ... that...'
- various more special constructions and contexts:
  'If there is an ... it is that...'
  'What is ... and ... is that...'
  'it seems/I find it/it may be ... that...'

Lastly, I carried out a few additional processing operations of roughly three types, in isolation or in succession:

- subsearches, typically on the one or two words preceding *that*;
- clean-up, including deletion of search keys like *here is*, etc
- alphabetical sorting, most revealing after clean-up, or 'fronting'
- select subconcordancing based on the syntactic criteria just mentioned, or on lexical lists ('it seems', 'I find/consider/... it').

As has been done by grammarians (recently in Greenbaum & Quirk 1990) for their own purposes, let me remind you of the synonymy between the following:

- what gets me/annoys me is that...
- what's most annoying is that...
- the annoying thing (about it) is that...
- there is the annoying fact that...
- it is annoying that...
- it is a nuisance that...

The lexicalizations focused on in this paper, in Adjective/Noun + reported-clause constructions, are the lexical end of a cline, with cleft constructions and the like at the other, syntactic end. (Note that French translations often resort to the pseudo-cleft construction.)

In English a syntactically and semantically interesting variant is a blend of adjective and noun in which both can have a reported clause;

this problem in itself might deserve detailed attention in another paper, so I will not dwell upon it here, and just give one or two examples:

(16) **clear documentation**
... there was 'clear documentation' that on at least four occasions Col. North was asking... (*Wall Street Journal*)
Cf 'It is clear that...' and 'There evidence/?documentation that...'
Cf ?'There is ample documentation that...'

(17) **encouraging signs**
There are encouraging signs that...
Cf 'It is encouraging that...' and 'there are signs that...'

Cf *jibe* defined as 'rude or insulting remark, in (15) above. Except for the words *evidence* or *proof* in its definitions, we have no clue, in OALDE, COBUILD, LDOCE, or BBI, for instance, that *documentation* can take a reported clause. Such blends seem only superficially similar to semantically equivalent collocations of the type BBI and many people in this audience have investigated (for *evidence*, a purple patch in BBI, no less than twenty-seven adjective collocates are entered, including *ample*, *cogent*, etc, but not *clear*). My claim then would be that such Adj/N+*that* combinations are not lexical collocations in the normal sense, especially if both the constituent adjective and noun can be used independently with a reported clause.

Computationally, I usually carried out three or more operations in succession: for example, selection of 'is/was/etc... that' contexts (which favour reporting adjectives), or '... is/was/etc that' contexts (which favour reporting nominals), some clean-up to remove irrelevant material, and lastly alphabetical sorts, or more concordancing.

Such displays of citations go to show that, in a corpus of this magnitude, we have no difficulty getting plenty of examples of the constructions most likely to favour the use of adjectives and nouns with reported clause, particularly 'presentative' existential *there* (Greenbaum and Quirk 1990: 428), and pseudo-cleft sentences, with a *that*-clause as complement and in which 'noun phrases of general reference' ('The reason... was that/because') commonly occur 'in place of the wh-item' (*op cit*: 414-415). In many such constructions, the underlying motivation may be to 'bring an entire proposition to the attention of the hearer', as the grammarians suggest in a revealing footnote (*op cit*: 424) on extraposed subjects of the type 'It has to be said that/It is a fact that', alternating and equivalent in this to clausal objects ('One finds that...').

Table 2: First ten lines of a file with some 150 citations (*The Independent*)

What is new is |that we have moved from too few qualified nurses...
(what is on offer in the market. Another Prais theme is |that...)
What is surely unacceptable is |that, had Mr Rushdie's work been...
What is amazing, therefore, is |that the agency that has the responsibility...
What is beyond doubt is |that after Mr Mandela is released, the question of...
What is baffling about H G Wells is |that he got his own way so effortlessly...
What is hard to dispute, however, is |that the Callaghan government in 1976...
What is perhaps more contradictory is |that putting legally married women...
What is so stimulating here is |that such questions are, so to speak,...
What is new is |that this hypothesis has now been tested for the first...

The vacillations about, and inconsistent treatment of, N+*that* and Adj+*that* in learners' dictionaries may in part be explained by the fact that there is a case for considering the two sets as somehow open-ended. This appears most clearly in constructions like 'What is Adj is that', 'Another N is that', where it is not always clear whether the item filling the slot accepts other constructions with a reported clause. Corpora, no matter how large, do not always provide crystal clear cases like the following; see *and* in (19) and (20):

(18) **crystal clear**
I'd like to make it crystal clear that I do not agree with these proposals (LDOCE, New edition; unlabelled)

(19) **ironic** *and* **indicative**
What is most ironic and perhaps most indicative is |that Mr. Reagan himself seems to have forsaken Reaganism (*Wall Street Journal*)

(20) **sceptical** *and* **cross**
... sceptical and a bit cross |that ...(*Wall Street Journal*)

Finally, I have appended extracts from my cumulative concordance of *that*-citations. Appendix 7 gives some contexts favouring reporting adjectives: a selection of citations of the type 'I find it Adj that ...' and one of examples of pseudo-cleft sentences. In Appendix 8, I had to restrict the selection to multi-word N+*that* citations (a ° sign identifies the content words which make up these semi-idiomatic expressions).

Appendices 9 (adjectives), and 10 (nouns) provide provisional lists of items not in my original checklists which were found to have at least one attested example with reported clause of at least one type. In the case of reporting nouns, I wish to emphasize that the existence of the structure 'a(n)/the N is that' does not necessarily imply that of the corresponding appositive structure:

(21) 'N is that' and appositive *that*-clauses

(a) The subtext of Mr Tebbit's thesis is that Mrs Thatcher is out of touch
Cf ?The subtext that she is out of touch
Cf ?? the text that ...

(b) A lovely thing is that ...
Cf ? the lovely thing that ...

In addition to the study of such syntactic problems, and of multiword nominals with reported clauses, further work now in progress includes the compilation, from our 120 MB cumulative corpus of citations, of those reporting nouns and adjectives that were not uncovered by the methods described in this paper. A provisional conclusion would be, however, that, in the absence of more reliable thesauri and more formalized dictionary definitions, grammatically defined search keys are the most productive ones.

Note

1. This paper was presented at the Thirteenth ICAME Conference, 3-7 June 1992, Nijmegen.

References

ACL/DCI, CD-ROM, Vol 1, 1991.

Benson M., E. Benson and Robert Ilson (eds). 1985. *The BBI combinatory dictionary of English*. Amsterdam: John Benjamins.

Greenbaum, Sidney and Randolph Quirk. 1990. *A student's grammar of the English language*. London: Longman.

Hornby, A.S. *et al.* (eds). 1974 and 1989. *Oxford advanced learner's dictionary of current English*. 3rd ed (1974), 4th ed (1989). Oxford: Oxford University Press.

The Independent on CD-ROM, 1989-90.

Procter, Paul *et al.* (eds). 1978. *Longman dictionary of contemporary English*. London: Longman.

Simpson, J.A. and E.S.C. Weiner (eds). 1989. *The Oxford English dictionary*. 2nd ed. Oxford: Clarendon Press.

Sinclair, John M. *et al.* (eds). 1987. *Collins COBUILD dictionary of the English language*. London: Collins.

Sinclair, John M. *et al.* (eds). 1989. *COBUILD English learner's dictionary*. London & Stuttgart: Collins & Klett.

Sinclair, John M. *et al.* 1990. *Collins COBUILD English grammar*. London: Collins.

Summers, Della *et al.* (eds). 1991. *Longman dictionary of contemporary English*. New ed. London: Longman.

Appendix 1: Checklist of adjectives

accepted acknowledged adamant admirable admitted advisable afraid
alarmed alarming amazed amused angry annoyed annoying anxious
anxious appalled appalling apparent apprehensive appropriate
arguable astonished astonishing astounded astounding avoidable
aware awful awkward bad believable best better bizarre
certain clear commendable compulsory conceivable confident
congruous conscious convinced correct credible criminal critical
crucial curious dangerous definite delighted delighting
deplorable depressed depressing desirable desirous despicable
disappointed disappointing disastrous disconcerted disconcerting
discouraged discouraging disgusted disgusting disheartened
disheartening distressed distressing disturbed disturbing
doubtful dreadful eager embarrassed embarrassing encouraged
encouraging essential established evident exasperated
exasperating excusable extraordinary fair fascinated fascinating
fated fearful fitting flattered flattering fortunate frightened
frightening frustrated frustrating funny furious glad good
grateful gratified gratifying great happy heartbroken honored
hopeful horrible horrified horrifying humiliated humiliating
imperative implicit important impossible improbable improper
inappropriate incomprehensible inconceivable incongruous
incredible indignant indubitable inevitable inexcusable
infuriating interested interesting intolerable ironic ironical
irrational irrelevant irritated irritating just known lamentable
likely logical lucky maddened maddening marvellous miraculous
monstruous mortified mortifying mystified mystifying natural
necessary nice normal notable obligatory obvious odd optimistic
outrageous overjoyed pathetic peculiar perplexing plain plausible
pleased pleasing possible predestined preferable preordained
probable proper proud puzzled puzzling queer rare reassured
reassuring regrettable relieved remarkable reported resolved
revolted revolting ridiculous right ruled rumored rumoured sad
said satisfied satisfying scandalous shameful shocked shocking
significant silly sorry strange sure surprised surprising
suspicious terrible terrified terrifying thankful thinkable

thrilled thrilling tickled tolerable tough tragic true
unavoidable unaware unbelievable uncanny understandable
understood undesirable unfair unfortunate unhappy unjust
unjustifiable unlikely unnatural unnecessary unthinkable untrue
unusual upset upsetting urgent usual vital weird well-known
willing wonderful worried worrying

Appendix 2: Checklist of nouns

acceptance accident accusation accusations acknowledgement
actuality adage admission advantage advice agreement allegation
analysis anger announcement answer anticipation anxiety appeal
approach argument arrangement aspect assertion assumption
assurance attitude attraction basis beauty belief benefit best
bet bit boast canard case catch cause caveat ceremony certainty
certitude chance chances change characteristic charge charm claim
cliché coincidence comment compensation complaint concept
conception concern conclusion concurrence condition confirmation
conjecture connection consciousness consensus consequence
contempt contention context convention conviction criticism curse
custom danger decision declaration decree deduction defect
defence demand demonstration denying design desire dictum
difference difficulty dilemna disadvantage discomfiture discovery
dissatisfaction distinction doctrine doubt dream effect entreaty
error essence estimate evidence example exception expectation
explanation fact factor faith fallacy fantasy fascination fault
fear fears feature feeling fiat foreboding forecast
generalisation glory gospel graveness grounds guarantee guess
hint hope hunch hypothesis idea illusion image implication
importance impossibility impression indication inducement
inference injunction insight interpretation intimation intuition
irony joke joy judgement justification key knowledge law legend
lesson likelihood line meaning merit message miracle
misapprehension misconception misfortune moral motion myth news
note notice notion objection observation obsession odds offer
omen omission opinion order orders originality outcome part
peculiarity penalty perception peril picture pities pity pleasure
point policy possibility prayer prediction premise premonition
presentiment presumption presupposition pretence pretext pride
principle probability problem promise pronouncement proof
prophecy proposal proposition prospect protestation providence
provision proviso range reaction reality reason reassurance
recognition recollection recommendation record reflection regret
relief remark reminder report request requirement resentment
resolution result revelation rider risk rub rule ruling rumbling
rumblings rumour saying scenario secret sensation sense shame

side sidelight sign significance snag speculation stipulation
story submission suggestion suggestions superstition supposition
suspicion symbolism task terror test testimony theme theory
thesis thing thought tip touch tradition tragedy triumph trouble
truism truth unbelief understanding undertaking upshot verdict
version view virtue vision wager warning weakness whisper will
wisdom wish wonder word worry worst wrinkle

Appendix 3: Regularities in the COBUILD definitions (alphabetically sorted)

(i) Adjective senses with reported-clause code

If people are agreed on something or if they are agreed that something is the case, they have reached a joint decision or a particular conclusion on it.
If someone is anxious to do something or anxious that something should happen, they very much want to do it or very much want it to happen.
If someone is positive about a fact or decision, they are very sure that it is true or correct and have no doubts.
If something is apparent to you, it is clear and obvious to you.
If you are afraid that an unpleasant or awkward situation or event will happen, you are worried about it and want to avoid it.
If you are aware of something or aware that something exists or is happening, you realize it because you hear it, see it, smell it, or feel it.
If you are aware that something such as an important problem or difficulty exists or if you are aware of it, you know about it, either because you have thought about it or because you have just noticed it.
If you are certain about something, you have no doubt in your mind about it.
If you are conscious of something, you notice or realize what is happening.
If you are conscious of something, you think about it more than other people do, because of the unusual or special way in which it affects you.

(ii) Noun senses with reported-clause code

A chance is the extent to which something is possible or likely to happen, especially something that is pleasant or desirable.
A claim is a statement that something is true or is a fact, although other people might dispute it and not believe it.
A comment is a statement of opinion about something.
A concept is an idea or abstract principle which relates to a particular subject or to a particular view of that subject.
A conception is a general idea that you have in your mind when you think about something.
A conclusion is something that you decide is true as a result of knowing that other things are true.
A conviction is a strong belief or opinion.

A declaration is a firm, emphatic statement which shows that you have no doubts about what you are saying.
A dream is a situation or event which you often think about because you would very much like it to happen, but which you know is probably not possible. See also pipe dream.
A guarantee is also a formal written statement of someone's intention to do something, or their acceptance of responsibility in a particular situation.
A guarantee is also a promise that you will do something, or that something will definitely happen.

Appendix 4: Checklist of adjectives selected from the COBUILD thesaurus

abhorrent acceptable adequate agreeable amazing ambiguous
appreciable appreciative ashamed aspirant awesome careful
charming comic comprehensible concerned confusing contemptible
content contented contrived controversial convenient convincing
cross debatable decent delightful desperate despondent
devastating disgraceful disloyal dispiriting disquieting
distrustful dubious eerie equitable excellent exceptional excited
expectant fantastic feasible fine firm fishy fit foolish
forgivable foul fundamental ghastly graceful guilty hapless
harsh hesitant hideous hilarious honourable hopeless horrific
horrifying ignorant immoral incensed inconsiderate indifferent
inequitable insufferable intriguing likeable ludicrous mad mean
miserable nasty naughty noble noteworthy noticeable objectionable
oblivious okay outraged perilous perplexed phenomenal pitiable
positive predictable preposterous questionable rational realistic
reasonable regretful relevant resolute safe sardonic scared scary
sceptical splendid striking stupid terrific tremendous tricky
typical unbearable uncomfortable uncommon uneasy unimaginable
unintelligible unnerving unpleasant unreasonable unsatisfactory
unseemly unsuitable vain vicious vile wary wicked wise

Appendix 5: Checklist of nouns from the COBUILD thesaurus

appreciation arrogance aspiration assessment audacity awareness
caution choice communication confusion critique delight delusion
desirability despair disgrace dishonour drawback eagerness
evaluation excuse extent fad failing falsehood fancy farce
fashion fate feat fiction finding fixation gag gamble gossip
grudge hallucination harmony hazard honesty honour horror
incentive inconvenience indictment inkling inspiration
instinct instruction intention interest legislation lie longing
marvel maxim misgiving mistake mystery norm notification
objective obscenity obstacle occurrence outlook penchant
perspective philosophy plan position posture prognostication

rationale rebuke regulation response responsibility reverie
reward ritual romance saga sarcasm satisfaction scare statement
tidings trust vista

Appendix 6: Adjectives from the COBUILD thesaurus found with reported clause

abhorrent acceptable amazing careful charming concerned content
contrived controversial convenient convincing cross debatable
delightful desperate despondent disgraceful disquieting dubious
excellent excited fantastic feasible firm forgivable fundamental
guilty hilarious ignorant immoral incensed intriguing ludicrous
mean noteworthy noticeable oblivious outraged perplexed positive
predictable preposterous questionable rational realistic
reasonable relevant resolute scared sceptical splendid striking
terrific tremendous typical unbearable uncomfortable uncommon
uneasy unimaginable unnerving unpleasant unreasonable
unsatisfactory unseemly

Appendix 7: Some contexts favouring reporting adjectives

(i) 'I find it _ that'

I find it abhorrent |that
I find it amusing |that
I find it absolutely baffling |that
I find it a little discomforting |that
I find it a little more impressive |that
I find it inconceiveable |that
I find it insulting |that
I find it pitiful |that
I find it quite startling |that

(ii) 'what is _ is that'

What is absurd is |that
What is agreed upon is |that
What is amusing, though, is |that (less _)
What is anticipated, however, is |that
What is baffling about H G Wells is |that
What is bewildering to market experts is |that
What is boring about it is |that (particularly _)
What is complex about this case is |that (so _)
What is confusing is |that (not _)
What is consistent beneath these shifting surfaces is |that
What is contradictory is |that (perhaps more _)
What is demanded is |that

What is different about this round is Ithat (significantly _)
What is difficult for Solidarity is Ithat
What is emerging is Ithat
What is exciting is Ithat (really _)
What is familiar is Ithat (less _)
What is galling to doctors and drug companies is Ithat
What is impressive about it is Ithat
What is indisputable is Ithat
What is inexplicable is Ithat
What is needed is Ithat
What is new about our theory is Ithat (really _)
What is newsworthy is Ithat
What is not cited by the Keynes bashers is Ithat
What is not in dispute is Ithat
What is not disputed is Ithat
What is not in doubt is Ithat
What is not mentioned is Ithat
What is overlooked in the article is Ithat
What is problematic about the Jersey City takeover is Ithat
What is problematic is Ithat
What is proposed is Ithat
What is relevant is Ithat
What is sick about the joke is Ithat
What is stimulating here is Ithat (so _)
What is stupid about Lichtenstein is Ithat (so _)
What is telling is Ithat (very _)
What is tiresome is Ithat
What is unacceptable is Ithat
What is unarguable is Ithat
What is unattractive is Ithat
What is undeniable, however, is Ithat
What is undisputed is Ithat
What is unique about Nadir is Ithat
What is unique is Ithat
What is unnerving about First Bank is Ithat
What is unusual about the Lehmann case is Ithat
What is well-known is Ithat (less _)
What is worrisome is Ithat (particularly _)
What is worse is Ithat
What is wrong is Ithat

Appendix 8: Some contexts favouring reporting nouns (restricted to examples of multi-word nominals in the context 'is that')

a less diplomatic °way of °putting it is Ithat

another *stumbling-block is that
another *stumbling *block is Ithat
another *way of *describing the situation is Ithat
another *way of *putting it is Ithat
his *ace *card is Ithat
my *gut *feel is Ithat
one *bright *spot for Compaq is Ithat
one *bright *spot in the outlook for housing is Ithat
one *bright *spot is Ithat
one *crumb of *comfort for the government is Ithat
one *loose *end is Ithat
one of its favourite *lines of *attack is Ithat
one possible *bright *spot is Ithat
one possible *stumbling *block to the acquisition is Ithat
one predicted *stumbling *block is Ithat
the biggest *stumbling *block is Ithat
the *bright *spot was Ithat
the *disturbing *element is Ithat
the *drawing *card is Ithat
the *end *game is Ithat
the *long and *short of it is Ithat
the main *conjuring *trick in ... is Ithat
the main *source of *frustration is Ithat
the *master *stroke is Ithat
the most *obvious, but most *unlikely, is Ithat
the *name of the *game is Ithat
the *nub of their *complaints is Ithat
the *old *song is Ithat
the one *bright *spot is Ithat
the one *bright *spot on ... is Ithat
the only *bright *spot for NATO, he says, is Ithat
the only *common *thread, analysts said, is Ithat
the only *saving *grace is Ithat
the only stated *common *ground is Ithat
the overriding *common *denominator is Ithat
the *prime *value of this short book is Ithat
the *quid *pro *quo is Ithat
the *rallying *cry for the Immigration Bill was Ithat
the *reverse of the *coin here is Ithat
the *rule of *thumb is Ithat
the *saving *grace is Ithat
the *saving *grace of the system is Ithat
the *saving *grace, until now, was Ithat
the *short of *it is Ithat
their one *strong *card is Ithat

Appendix 9: Reporting adjectives not in the original checklist which were found in the corpus

abhorrent absurd acceptable agreed amazing amusing anticipated
ashamed baffling careful charming concerned content
controversial convenient convincing cross debatable disgraceful
disquieting dubious emerging excited exciting fantastic feasible
fine fit forgivable fundamental galling ghastly guilty harsh
hilarious ignorant immoral impressive incensed inconceiveable
indisputable inexplicable insulting intriguing ludicrous mad
noteworthy noticeable oblivious outraged perplexed positive
predictable preposterous problematic proposed questionable
rational realistic reasonable relevant resolute scared scary
sceptical sick splendid startling striking stupid telling
terrific tremendous typical unacceptable unbearable
uncomfortable uncommon undeniable undisputed uneasy unimaginable
unintelligible unnerving unreasonable unsatisfactory unseemly
wary worse wrong

Appendix 10: Reporting nouns not in the original checklist which were found in the corpus (including a few adjectives used as head nouns)

abuses account achievement aim alternative analogy anomaly
apprehension assessment asset attribute attributes awareness
axiom backdrop background bad bait (plea) bargain beauties beef
bent bias blessing bonus boost breakthrough brilliance burden
buzz byproduct calculation card carrot (note of) caution
certainties challenge chat circumstance clause cleverness cliche
clincher clue coda comfort comparison complication compliment
conceit concensus concession confusion consolation construction
contract contradiction conundrum core corollary cover crime
crisis criteria crux curiosity damage damnation deal defense
definition delight delusion detail development difficulties
dilemma drawback drawbacks element encouragement ethos event evil
excuse experience extreme failure fallout figure finding findings
format frustration frustrations gamble game genius gist glories
gloss goal greatness gripe guidance handicap hazards headache
heart hindsight hook humiliation hurdle illnesses
impact impediment impetus import imputation incentive indicator
influence information injustice innovation inscription instinct
insult intelligence intent intention interest interests ironies
irritation issue item judgment lament legacy leitmotif lie
limitation linchpin link logic lure (no) matter maxim measure
mechanism memory mentality metaphor mismatch mistake mitigation

model mood motivation motive mystery nature negative nightmare
norm novelty object objective obstacle oddities oddity offence
offense one option other outlook paradox paradoxes pattern
payoff peeve perk perspective philosophy phrase pitch pitfall
plan platform plea plus poignancy position precedent precepts
preference preoccupation price procedure process product
prognosis projection protocol psychology purpose puzzle
qualification qualifications qualifier quandary quarrel quibble
quip rationale rationalization reading realities realpolitik
refrain rejoinder relevance remedy reply reputation requisite
reservation response restriction retort reward rider (golden)
rule rumor sadness safeguard saw scandal scare school (of
thought) scoop second sentiment settlement shift shock shocker
similarity sin situation solace solution source specter spectre
spirit stage stance statement statistic stereotype sting stories
strategy straw strength strengths structure struggle subtext
subvariant success surprise surprises tale talk target teaching
tenet tenets theories thinking third thread threat threats
thrust tone tragedies trait trend tribute trick twist utility
value vice way worries yarn

Corpus data processing with Lexa

Raymond Hickey
University of Munich

Abstract: The present article offers an introduction to the software system *Lexa* which has been designed to facilitate the processing of corpus data. The main applications of the system, such as lexical analysis or information retrieval, are discussed with typical cases being examined. After a brief explanation of what files types can be handled by the *Lexa* suite the question of text categorization is looked at. Then a detailed presentation of automatic tagging is offered. Particular attention is given to the degree to which such operations can be customized to users' needs along with the transfer of textual data to a database environment for the purpose of constructing lexical databases. The article concludes with a selection of further applications of the programme suite in the general field of corpus data processing.

1. Introduction

1.1. Purpose and scope

The purpose of the present article is to introduce the software suite *Lexa* to the community of linguists interested in learning about software for the management and processing of text corpora on a personal computer.[1] The system *Lexa* is a complete text retrieval system with its major emphasis in the general area of corpus processing, particularly the tagging and analysis of texts and the derivation of lexical databases from such texts and their subsequent handling with appropriate database management software. Given the scope of this introductory article only a brief sketch of some typical applications of the software can be offered. I have chosen to look more closely at the area of lexical and grammatical analysis of texts and to follow this with some references to databases, information retrieval and the processing of historical corpus

texts. Hopefully the descriptions below will convey to readers (and potential users) an impression of what the aim and scope of the *Lexa* suite is.

1.2. Availability of Lexa

The present suite of programmes consists of more than 60 executable files comprising some 4MB along with additional sample data. The set is self-installing and requires no particular hardware apart from a fixed disk with at least 5MB of free space and of course additional space for any primary corpus data which users may wish to process. Accompanying the software are 3 volumes (each between 250 and 300 pages in length) which contain both extensive documentation of the programmes and exemplary discussions of typical processing tasks. The volumes refer to typical data processing areas covered by the software, namely (i) *lexical analysis and information retrieval*, (ii) *database and corpus management* and (iii) *general file management*. The texts are intended to be suitable for beginners and include comprehensive glossaries of all technical terms used in the body of each volume. The programmes and texts have been published by the University of Bergen and are available from the Norwegian Computing Centre for the Humanities in Bergen as of Spring 1993.[2] As the software is intended for bona fide scholarly research there is no special copyright agreement concerning its use, nor is there any kind of programme protection.

1.3. Making use of Lexa

For computer users who are acquainted with the basics of personal computing the use of the *Lexa* suite should present no difficulties. It is organized as a collection of over 60 programmes.[3] Of these some are major and other are minor. To start with, the set can be surveyed by means of a so-called control centre. This is a programme which offers the user a brief summary of each member of the suite and allows him or her to load any programme, automatically returning to the control centre for renewed starting of a further programme. By these means the user can very quickly ascertain what the individual programmes of the suite actually do. An alternative launching pad for all programmes is available as a desktop which complies in its design to the SAA (*system application architecture*) standard of IBM which users will be familiar with from such environments as *Microsoft Windows*. Indeed, all major programmes employ a system of picklists available on an entry level to the particular programme, allowing the user to activate

any option of the programme by simply moving a highlight bar and pressing the Return key. Again for all major programmes, online help and mouse support are included.

Furthermore, configuration information is stored to disk and can be used during a later work session. As a matter of principle, all the main programmes can be run interactively or in the so-called *batch mode* in which a programme loads itself, gleans its configuration information from a setup file, executes and returns the user to DOS automatically without it being necessary to supply user input during the execution of the programme. The advantage of this is that various tasks can be executed automatically as a group without the user necessarily being present. The time factor involved in complicated and intricate processing tasks then becomes irrelevant. All programmes which collect information about texts or databases during their operation can write this to an output file (for later inspection with a text editor such as that supplied with the *Lexa* suite) apart from displaying information collected on the screen. Note that all input files for processing must be either ASCII texts or databases in the *dBASE* format for the *Lexa* programmes to accept them as valid input. This is not a restriction but rather a gain in flexibility over word processor files (such as those generated by *WordPerfect* or *Microsoft Word*) as the source files can come from any computer environment, not just a personal computer, e.g. from a mainframe or a Unix work station.

2. *Corpus data processing*

It should be mentioned at the outset of this section that the *Lexa* suite was designed to be used with any text corpus. The programmes make no assumptions about the source of input texts apart from their being pure ASCII texts. Nonetheless, users will notice that many references are made both within the documentation and with the software to the Helsinki Corpus of historical English texts (Kytö, 1991). There are definite reasons for this, which have to do with the association of the present author with colleagues in the Department of English in Helsinki, notably with Matti Rissanen and Merja Kytö, both of whom have been instrumental in realizing the Helsinki Corpus (Kytö and Rissanen, 1992:7ff.). I would be pleased to be mentioned in connection with the latter corpus and for my software to be used with it for data processing tasks. At this stage my only desire is to emphasize that the *Lexa* suite can be applied to any corpus, including the corpus of Irish English

being presently compiled by the present author or already available corpora, such as the Lancaster-Oslo/Bergen Corpus.[4]

2.1. Categorization of texts

All the programmes of the *Lexa* suite which process data can take as their input text files which are specified by the user. There are a variety of means for specifying such files. The easiest of all is for the user to select a file from a directory listing presented on the desktop of one of the data processing programmes. Another means is for users to enter a file template which encompasses the files to be affected by an operation to be performed. Such means are mechanical and depend entirely on file grouping according to the names used by the operating system. A more flexible system is available for all the programmes which perform information retrieval tasks. Here users can specify that a programme use as its input those files which are deposited in a so-called *list file*. The latter is a small ASCII file which consists of several file names, each on a separate line of the file. There need be no similarity in name between the files listed, this freeing one from the straightjacket of file names on the operating system level. The scope of this option is greatly increased if one considers carefully how such list files can be generated. To begin this discussion, allow me to present briefly what is known as a file header and the widespread format used for this, the *Cocoa file header format*.[5]

Among the text corpora available today many make use of a format for including information relating to the contents of files. A commonly used format is that called the Cocoa format which consists of a series of parameters which characterize the text in question.

1: <B = 'name of text file'>	2: <Q = 'text identifier'>
3: <N = 'name of text'>	4: <A = 'author'>
5: <C = 'part of corpus'>	6: <O = 'date of original'>
7: <M = 'date of manuscript'>	8: <K = 'contemporaneity'>
9: <D = 'dialect'>	10: <V = 'verse' or 'prose'>
11: <T = 'text type'>	12: <G = 'relation to foreign original'>
13: <F = 'foreign original'>	14: <W = 'relation to spoken language'>
15: <X = 'sex of author'>	16: <Y = 'age of author'>
17: <H = 'social rank of author'>	18: <U = 'audience description'>
19: <E = 'participant relation'>	20: <J = 'interaction'>
21: <I = 'setting'>	22: <Z = 'prototypical text category'>
23: <S = 'sample'>	24: <P = 'page'>
25: <L = 'line'>	26: <R = 'record'>

This information can be accessed by the information retrieval software

of the *Lexa* suite in the following way. A programme (called *Cocoa*) extracts the header information from any set of input files and deposits this in a database. Then, with the database manager *DbStat* one can load the database just created and impose a filter on it by which only those records remain visible which meet a certain condition. Assuming that one generates a database of the Cocoa header information in the files of the Helsinki Corpus, then one could specify a filter to which only those records (i.e. file headers) correspond which represent translations (Item 13) of Middle English (Item 6) prose (Item 10) texts. A list of the files for which this header information obtains can be generated by creating a list from the field information for Item 1 (*name of text file*). The list file created by these steps can in its turn be used as the source of the file names for an information retrieval operation with other parts of the *Lexa* suite so that only Middle English prose translations from the corpus are examined. In addition the user can specify with the retrieval programmes from the set (such as *Lexa Pat* and *Lexa Context*) that the Cocoa information of the files examined be enclosed in the output file of statistics generated during a search.

The example just given is typical inasmuch as it illustrates how different parts of the *Lexa* suite link up together. For any prospective users of the programme package it is essential to grasp the interrelationships between items of software. A disconcerting and sadly not uncommon experience of the present author is that users complain that some feature is not present when in fact it is, but they have not realised it as they fail to grasp the potential of certain programmes.

2.2. Lexical and grammatical analysis of texts

The following section is intended to convey an impression of what tasks can be accomplished by using the main programme of the *Lexa* suite. To begin with, a word of explanation regarding nomenclature is necessary. In the *Lexa* suite the main programme for carrying out lexical and grammatical analysis is itself termed *Lexa*. All other programmes consist of *Lexa* and a further word which refers to what function they perform. Thus the pattern matcher is called *Lexa Pat*, the programme for locating syntactic contexts is called *Lexa Context*, etc. The names of these files on the operating system level consist of the function word (or an abbreviation of this) preceded by the letter 'l', e.g. lpat, lcontext, etc.

2.2.1. Preamble: What is meant by 'text'

It is fair to say that any data which users of the *Lexa* suite will process

will initially be in text form, i.e. the files are in the so-called ASCII format. An exception to this is the special case where one commences with texts which have been indexed for use with particular software in advance (this is the case of the Helsinki Corpus which is available on CD-ROM in a pre-indexed form for use with the commercial text retrieval system *WordCruncher*). But even in such instances, the actual text files usually remain in the original ASCII format, i.e. they do not contain any information which is specific to a certain word processor. This is in sharp contrast to the situation with the text files one may generate with one's word processor on a personal computer. Here the file which contains a text will also include information for the formatted output of the text on a printer, e.g. information concerning the layout of the page (page length, left/right margins, etc.), and the attributes used for certain letters or words (boldface, italics, etc.). Formatting information is always specific to a particular word processor and so cannot be intelligently interpreted by some other programme. To analyse texts with *Lexa*, which have been processed or created with a word processor, these must be stored to disk without any formatting information (this option will always be available with one's word processing software).

Users of computers should thus bear in mind that in computing the term *text* has a very definite meaning. A text is a collection of informational units (usually bytes) which are arranged as an unstructured number of lines. There may well be a semantic structure to the text (determined by its contents) but for the computer a text contains no inherent structure. In computing, the term 'text' is frequently used somewhat loosely to refer to an ASCII (i.e. non-formatted) text.

With an ASCII text there is a pre-defined set of characters which trigger the end of a line: ASCII $13 and ASCII $10 (called *carriage return* and *line feed* respectively). Any programme processing an ASCII text thus knows where a line comes to an end.

2.2.2. Tagging a text corpus

Before any kind of useful lexical and grammatical analysis can be performed on a text corpus it is necessary for it to be tagged.[6] This is a task which may well have been carried out in advance by the compilers/distributors of the corpus. However it is not always the case. For instance the Lancaster-Oslo/Bergen Corpus is available in a pre-tagged form whereas the Helsinki Corpus is not. Thus, those users of the latter who wish to tag it (to what extent is a secondary matter) will require software such as *Lexa*. Note that the tagging scheme used for the Lancaster-Oslo/Bergen Corpus can be applied to the Helsinki Corpus in

either the original or a user-defined form (the decisions on what forms in a text are to obtain what grammatical tag from a set of tags are made by the user; the details of this procedure are outlined below). In essence, tagging works as follows: each word in a file is examined and a label is added to it to identify it grammatically.

Stretch of text before tagging

A marchant was ther, with a forked berd,
In motlee, and hye on horse he sat,
Upon his heed a Flaundryssh bevere hat,
His bootes chasped faire and fetisly

Stretch of text after tagging

A_ART marchant_NOUN was_VERB ther_PREP, with_ADV a_ART forked_ADJ berd_NOUN,
In_PREP motlee_NOUN, and_CONJ hye_ADJ on_PREP horse_NOUN he_PERPRO sat_VERB,
Upon_PREP his_POSSPRO heed_NOUN a_ART Flaundryssh_ADJ bevere_NOUN hat_NOUN,
His_POSSPRO bootes_NOUN chasped_VERB faire_ADJ and_CONJ fetisly_ADV

It is obvious from the above illustration that the tags are placed after the words they refer to and are separated by a single underscore (_, the character used can be specified by the user). This is a widely accepted convention (cf the London-Lund or the Lancaster-Oslo/Bergen corpora). The tag itself is the capitalised abbreviation used to unambiguously classify the word in question. Needless to say, for an ensuing grammatical analysis of any sophistication, it would be necessary to devise more refined categories than those used for illustrative purposes above.

Tagging may be done manually by the computer operator deciding as he or she goes through the text how each word is to be classified. However, the task is impracticable unless one resorts at least to a degree of automatic tagging. Both methods are available with *Lexa*, as is a combination of the two.

2.2.3. Lemmatisation

The term *lemma* is used in lexical data processing with the equivalent meaning of *lexeme* in general linguistics. A lemma is thus an abstraction of the set of inflected forms which are united by a common semantic core. For instance the attested forms *walk, walks, walked* and *walking* all belong to the lemma WALK. In the lexical analysis of a corpus,

the concern is then to group together inflected forms and somehow mark them so that their semantic affiliation is obvious.

2.2.4. Automatic tagging

2.2.4.1. Lexical tagging

With any type of computer analysis involving automatic procedures it is necessary for the computer operator to initially lay down the criteria which the system is to use for classification. When tagging a corpus with *Lexa* this can be achieved as follows. The computer operator creates a file with a list of tags contained in it. After each tag are listed those forms which can be given the tag in question. When involved in lexical tagging, the user enters the keyword *Lemma* on a line and after this the lexeme to which the ensuing form belongs.

```
LEMMA: SING
TOKENS: WORDS
sing
sings
singing
sang
LEMMA: WALK
TOKENS: WORDS
walk
walks,    etc.
```

Before initiating a tagging session the computer operator specifies which tag-list file is to be used for the run. The system reads the file and fills an internal table with the definitions contained in the tag-list file.

Technically the steps are as follows: a two-dimensional array is allocated in memory from the heap (that section of system memory which can be used for data by the programme which is currently running). One dimension of the array contains the names of the tags which are defined in the tag-list file; the second dimension contains the forms which are defined as being tokens of a particular tag. You can envisage this as a series of rows and columns with types and tokens occupying vertical and horizontal positions respectively. For every word in a text which is tagged, *Lexa* combs through the entire array of tags to see if the current word is a token of some tag or other. If the search for a match is positive the relevant tag is attached to the current word and the system proceeds to the next word.

For the *Lexa* programme a tag can be of two basic types: (i) it refers to a lemma, i.e. a dictionary entry which subsumes a whole series of inflected forms, in which case the tag begins with the keyword LEMMA or (ii) it indicates a word class or morphological category, in which case the keyword CLASS is to be found after the tag.

As can be seen from the above example, on the next line after the lemma the keyword TOKENS occurs; immediately following this is either the word WORDS or STRINGS. This is noted by *Lexa*, and when lemmatising a text the tokens which are found in the input file are then either treated as whole words or as strings. Let us take an example to see what advantage is to be gained from this. Say you have defined a set of prepositions as follows:

```
CLASS: PREP
TOKENS: WORDS
for
in
out
on
```

When later combing through a text, *Lexa* will only mark occurrences of these tokens as instantiations of the word class PREP (the same would apply to a lemma) which form whole words, thus avoiding incorrect tagging such as foreigner_PREP, intake_PREP, outgoing_PREP, button_PREP.

2.2.4.2. Grammatical tagging

When dealing with inflected forms it is no longer sufficient to use a list of lexemes as a basis for successful tagging. The solution is to create a tag-list file which consists of sub-word morphemes and to allow the data processing software to determine class affiliation on the basis of a morpheme being present in a word form or not. Consider the following extract from a list file for grammatical tagging of a group of Middle English texts:

```
CLASS: PAST_PART
TOKENS: STRINGS
#y*e#
#y*en#
CLASS: ADV
TOKENS: STRINGS
ly#
CLASS: PREFIXVERB
```

TOKENS: STRINGS
#pre
#fore
CLASS: FRENCHVERB
TOKENS: STRINGS
ceive#

It should be obvious just what type of returns one is expecting with such a list. Note that the *fore* tag for a prefix verb will not of course yield unambiguous results, as words like *forehead, forelimb* would be returned if present in a text which is examined. Equally, the ending *ly* will return words like *fly* which must be re-classified manually afterwards.

One solution to the difficulty of unacceptable returns is to perform some other type of tagging beforehand which would capture these forms. Once they are tagged they will not be re-tagged by the system.

Another solution would be to create a stop word file (see below) with those forms listed in it which one does not want tagged. Of course this alternative is really only viable if the set of potentially undesirable tags is fairly small.

With the *Lexa* programme, if the symbol '#' (or a user-specified word delimiter) is placed before an affix, then it must occur at the beginning of a word; if it is placed at the end, then it must be at the end of a word. The word delimiter can be used at the beginning *and* end of a token. DOS wild cards, * and ?, can also be used to leave (one or several) characters unspecified.

2.2.5. Cumulative tagging

Not all the words of a text need to be tagged on one run. In fact it is sensible to tag the most obvious words (those which constitute a small closed class, of items) first and then gradually work on to the more difficult classes with hopefully only a small number of non-classified items left which have to be dealt with manually by the computer operator at the end.

When the data processing software examines a text, a mechanism can be used to determine whether any given word which it strikes upon has already been tagged or not. If every tag begins with a pre-defined character, say an underscore, and if the underscore does not occur as a constituent of any normal word of the text, then any given word can be examined to see if it contains an underscore. Should this be the case, then the word has been tagged on a previous run; if not, then the data processing software is to attempt tagging this time.

2.2.6. Manual tagging

No matter how good the tagging algorithm is, there will always be a residue of word forms which cannot be automatically classified. These must be tagged manually. To do this, the programme must demand that the computer operator decide on the tag to be attached to any words found in a text being examined which have not yet been classified. Bear in mind that manual tagging is always necessary with ambiguous forms, as the data processing software can only use formal criteria to determine class affiliation. Within *Lexa* there is a text editing level with special features which refer to tagging. The text currently loaded in memory can be viewed and edited at any stage (from the desktop). When editing a text, manual tagging can also be carried out by means of a number of inbuilt macro facilities.

2.2.7. Stop words

The easiest forms to tag are those which form a small closed class, e.g. the articles in English. However, these forms are usually of little interest to the linguist examining a corpus. What is then desirable is to filter them out and concentrate on the remaining forms. This can be achieved quite easily. The first step to this end is to create a list of those words (called *stop words* in computer jargon) which are to be ignored. When the data processing software then examines a text file, it first checks to see if a given form has been labelled as a stop word (by looking it up in an internal table). If so, the form is ignored and it precedes to the next.

Evading stop words can be achieved either by excluding them from a tagging operation or by erasing them from a file to start with. One might care to create a temporary version of a text file without stop words as this would speed up tagging later (after all there would then be no cases in which forms are examined and then discarded by the system).

2.2.8. Locating and altering tagged forms

At any point when processing a corpus, it may be expedient to both tag certain sets of forms and then locate them to see just what words were affected by the tagging. This can be realized within *Lexa* when dealing with single files. If a broader scope is required, covering a group of files, for instance, then the easiest way of satisfactorily locating tagged forms involves using one of the many information retrieval programmes in the *Lexa* set. The supplementary programmes one can

avail of are *Lexa Pat*, *Lexa Search* or *Lexa Context*, for text files, or *Lexa DbPat*, for databases.

It may well occur that, once one has tagged a text or some texts, one wishes to alter the tagging done. There is a general utility *Lexa Sweep* which can be used, among other things, for this purpose. One specifies the form of an old tag, that of the new one to replace it and the set of files to be affected by the operation. One can also use *Lexa Sweep* to remove tagging, i.e. one says what tag is to be located and leaves the replace string empty. This removes a tag without inserting a substitute in its place.

2.2.9. Multiple tag files and input texts

When *Lexa* is run in the so-called *interactive mode* the user chooses an input text from a directory listing offered on the *Lexa* desktop. By choosing a further option from the relevant picklist one can then proceed to tag the text chosen. This procedure is sensible when one is getting acquainted with computerized tagging and the functioning of the programme *Lexa*. However, with time one will wish to be more flexible in data processing. To achieve this, *Lexa* must be executed in the *batch mode*. By this is meant that all the information necessary for the operation is specified in an initialization file. The programme then derives the values for all its user-specifiable parameters by examining this file on loading. One can demand that *Lexa* analyse a series of texts by using a file template (a specification with one or both of the MS-DOS wild cards * and ?) instead of an explicit file name as input text for analysis. *Lexa* will then examine any files found in the data directory which match this template. The same technique can be applied when specifying the name of the tag list file to be used. Should a file template be entered at this point in the initialization file, then *Lexa* will attempt tagging each file of the input text template with tag definitions from each of the files in the tag file list template.

During batch mode operation, *Lexa* informs you of what it is doing. However, no user input is necessary so that the presence of the user is not required. Furthermore, very large files can be processed in the batch mode. Should these not fit into available system memory, then *Lexa* can use the so-called *file-slice mode* in which it loads a section of the text currently being examined and, when finished, proceeds to the next section until the file has been analysed completely.

2.3. Constructing lexical databases

A frequent desideratum when lexically analysing a corpus is to construct a dictionary with grammatical information included on the word forms which constitute the dictionary. Such a task becomes quite easy with a lemmatised text. The first step (or rather sequence of steps) is the complete lemmatisation of the texts in question. Once this has been achieved the data processing software can now extract information from the text and deposit it in a database. Recall that a database is a structured file which consists of a number of records, each in turn consisting of a number of fields. A non-electronic parallel would be a box of index cards. Each card corresponds to a record, and assuming that there are ordered divisions on each card, then these would represent the equivalents of record fields. A typical lexical database has one record per word form.

The programme *Lexa* constructs a (primary) lexical database by generating an empty database with a minimum of four fields as follows (this is all that is required at this stage; lexical databases can of course be manipulated later with the database manager of the *Lexa* suite, *DbStat*).

Field 1: TOKEN
Field 2: LEMMA
Field 3: FREQUENCY
Field 4: REVERSE

Each word form in the database occupies a record of its own. The system starts by checking to see if a particular record is already present in the database. If not, a new record is appended and the word form is entered automatically in the field TOKEN. The lemma is extracted from the tagged word by locating the lemma divider character (by default an underscore) and copying the remainder of the word form (up to the next space or item of punctuation) into the field LEMMA. The field FREQUENCY is incremented each time an occurrence of the particular type of that record is found. Lastly, the field REVERSE contains the word form in reverse order. The idea behind this is to allow users to create a reverse dictionary (by sorting the database on the field REVERSE), thus making it much easier to recognize what inflectional information is contained in the word forms of the database.

After the process has been completed, you are left with a database which has as many records as there are unique word forms in the corpus examined. Note that, should a word form not be lemmatised in

the corpus for some reason, then the form is nonetheless added while the field LEMMA is left empty.

Apart from the database type just outlined above, it is possible with *Lexa* to generate a database which has one record per lemma. This is a secondary database, which is realised by first creating an empty database manually (e.g. by deriving a shell database from a *Lexa* database via the appropriate option on the *DbStat* desktop) and then importing the information from a *frequency list file* into it subsequently. The information in the latter type of text file (which is generated by an appropriate option in *Lexa*) is organized into lines with three items on each: a *unique* word form, the lemma attached to it and the number of times it occurs in the database (frequency). These items of information always begin in the 1st, 33rd and 49th columns of each line in the text file respectively. Due to this organization it is easy to import the information into a database by treating the frequency text file as an SDF (= *system data format*) file and using it as the source of a text importation operation with a database manager (such as *DbStat*). The databases generated by *Lexa* are always in the *dBASE* format. This is by far the most commonly available and readable format on personal computers. The resulting lexical databases can be read by virtually any other database manager in addition to the one supplied with the *Lexa* suite.

2.3.1. Generating database-readable text files

In order to move the data of a text file to a database environment it is essential to either create a database or a file which can be read directly by a database. The latter course of action is covered by an option within *Lexa*. It generates a so-called delimited text file from the text in memory. Using a specially reserved character as a delimiter of certain contents on each line, an output text is created which can be read by a database manager and which leads to the information in the text file being properly assigned to the fields of a database.

2.3.2. Merging textual information with databases

As a corpus will in all probability consist of a number of text files, generating a database from the word forms of an entire corpus will require that the data from each text file be transferred to a database. However, it would be pointless to create a new database each time a text is analysed. Instead, what one needs is an option in which data is added in a cumulative fashion to a single database so that it reflects the lexical structure of more than one input file.

This is realised with a further option within *Lexa*. For the first text to be analysed one creates a database with the *Generate database* option. With all subsequent files one merges databases. In doing so, one must first of all choose a database to merge textual data with. Care should be exercised here that the database chosen is one which was generated by *Lexa* at some previous stage. If not, *Lexa* issues an error message and refuses to continue.

Assuming that the database is acceptable to *Lexa*, it now combs through it and undertakes one of two steps: (i) adds the word form in the current text file in memory to the external database if this form is not already contained in the latter, (ii) increments the frequency field of the database, should the current word form from the memory text already occur in the external database.

2.3.3. Statistical operations and databases

The database manager of the *Lexa* suite is especially geared towards the processing of numerical data. To this end it contains a wide range of statistical options. These can be applied to the frequency figures generated by many other programmes, such as the main programme *Lexa*. All such programmes can place the result of some operation which generates frequency tables in a text file of a special kind which can then be read into the field of a set of records with the database manager.

The statistical options available with *DbStat* fall into three main groups:

(i) Options for preparing or arranging data.
(ii) Options for determining central tendency.
(iii) Tests which involve (two) sets of data.

The first group will perform such tasks as ranking data, sorting them, generating interval and frequency lists or displaying the range in a set of data. In this case, as with others connected with calculations with *DbStat*, a set of numerical data is defined by the entries in a numerical field for the records of a database.

With the second group one has a series of options which determine central tendency with a set of data. Examples of these are median, mode, interquartile, variance, standard deviation (biased and unbiased) apart from simpler types of calculations.

The purpose of the third set is to carry out operations which are particularly suited to the type of non-parametric data found in linguistic material. Note that for inferential statistics two sets of data are always required. Three possible relationships may obtain between these:

(i) One set may represent a set of expected values and the other a set of observed values. (Chi-square)
(ii) The two sets are possibly correlated. (Pearson, Spearman)
(iii) One set may be a sample and the other the parent population from which the sample is putatively drawn. (Mann-Whitney, Wilcoxon, Sign-test, F-test)

The types of test available in *DbStat* for the particular set of data are indicated in brackets above.

It would go far beyond the scope of the present introductory article to explain and illustrate the statistical options which are put at the user's disposal with the database manager *DbStat*.[7] It must suffice at this point to hint at them; I cannot do anything else but refer the interested reader to the documentation accompanying the software which contains greater detail on the use of such options.

2.4. Generating concordances

A further set of features in *Lexa* is concerned with the generation of text files in which word forms are highlighted in order to easily recognize the context in which they occur. These are traditional types of files to be found with concordance programmes and are included at this point to offer similar facilities to users of *Lexa*.

Concordance file (i). This option generates a so-called KWIC concordance. The abbreviation stands for "key word in context" and, as the name implies, each occurrence of unique word forms is highlighted (by spacing on either side) in the context in which it is to be found.

Concordance file (ii). The second type of concordance file has similarities with what is known as a KWOC file, from the designation "key word out of context". The keyword is, however, not so much out of context as not centred in the line in which it occurs. This type of file simply contains the tokens of word types enclosed in curly brackets for easy recognition.

Concordance files normally contain all the unique forms found in the text file currently in memory. However, if you set the relevant option in the initialization file to 'on', you can force *Lexa* to create a concordance file with only a selection of word forms. These are contained in a text file which is also specified in the initialization file. A word list for concordance generation consists of a number of words, each occupying exactly one line in the input text file. This option can be tested with the supplied file *excerpt.frm*.

2.5. Lexical density

Token lexical density. A text file is created with the present option which contains the unique word forms of any text arranged in ascending alphabetical order of their frequency, offering the user a picture of the density of word forms in the text.

Lemma lexical density. This is similar to the previous option with the difference that the lemmas of the word forms (i.e. the tags) in a chosen text are arranged in the output text file according to frequency of occurrence.

3. Information retrieval with Lexa

One very large area, which has only been touched on indirectly so far, concerns information retrieval. By this is meant the selective extraction of user-specified information from the texts of a corpus. Note that *Lexa* can handle such tasks with both texts (the normal form of a corpus) and databases (a derived form).

The main retrieval programme is *Lexa Pat.* Its basic function is to locate user-specified strings in text(s), writing the results of a search along with statistics gathered during such a search in a text file. The programme contains a number of additional extras to improve flexibility. For one thing, sets of files can be combed through. To indicate this one can use a DOS file template or a list file, as discussed above (see 2.1.), which can be generated by means of the programme *Cocoa,* thus restricting searches to a (user-specified) subset of files. Furthermore, the forms searched for can be indicated by a normal file template or by the user conveying the name of a list file in which a series of words which are to be searched for are included. The forms in such a list file may in their turn also include DOS wild cards to broaden the base of possible matches which might be returned by the system.

As with the major programmes of the suite, *Lexa Pat* is configurable, writes all the information which it collects during its operation to a text file and, most importantly, can be run in the so-called batch mode (again, see above 1.3.).

3.1. Locating syntactic contexts

The information retrieval software of the *Lexa* suite is not confined to the location of single word forms. Very often the linguist will be interested in finding syntactic contexts. The programme *Lexa Context* is intended to fulfill this need. Basically, what the programme will do

is to look for any word or string and then locate a second word or string within a specified number of words or characters, thus returning a syntactic context. The only requirement for the programme is that the context be formally specifiable. The user can use the DOS wild cards ? (for one unspecified character) and * (for more than one such character) in the words and/or strings used in a search, e.g. locate contexts within the following frame: *that*, up to 8 intervening words, **ed*. This would in all probability return contexts of relative clauses which end in past forms of verbs. Of course, the reliability of the returns depends on how well the context can be and is in fact specified by the user. Nonetheless, few contexts will be entirely unambiguous. A solution to this quandary is to allow the user to decide whether the context returned by the programme represents a genuine find for the context the user is looking for. You can force *Lexa Context* to display each context on screen and to ask the user whether it is genuine or not. By these means the user can decide what contexts are acceptable and hence to be added to the statistics which are collected during a search.

This programme has been used effectively by the present author to look at the syntactically deviant forms in the dramas of John Millington Synge[8] (part of the corpus of Irish English under preparation). It was successfully employed to locate structures like *after* + present participle as the indicator of a perfective aspect in Irish English and the use of *for to* + infinitive in clauses of intention as well as general fronting with cleft sentences introduced by *it is* and a topicalized element from a sentence.[9] Note that *Lexa Context* can take the sentence as its primary unit of investigation. The user conveys to the programme what items of punctuation signal the end of a sentence. Going on these, it divides the text it examines into sentences and returns statistics which refer to this organizational unit.

4. Additional facilities

4.1. Normalization of texts

An editorial task which arises quite frequently is the normalization of texts. There are a variety of reasons why this should be so. A common one is to reduce the distracting effect which irrelevant data can have on users analysing a text or set of texts. Such normalization might involve the levelling of irregular verb forms with a text which one is investigating for some other information than verb composition. This task can be achieved easily with a utility in the *Lexa* suite called

DbTrans. Essentially what it does is to examine an input text or texts, and going on a dictionary database which is conveyed to it by the user carries out a series of substitutions. In the hypothetical example just quoted, the user would specify what variant verb forms are to be regarded as manifestations of what normalized forms. The programme then consults the database specified and, if it locates a form in the input field of the database, replaces this by that in the output field. The net result is a group of replacements which, if the substitutions are correctly specified by the user in the database consulted, leads to normalized output text.

4.2. Display of older texts from the Helsinki Corpus

In the compilation of the Helsinki Corpus its designers made a wise decision to encode special symbols which are necessary for Old and Middle English by using so-called 'escape sequences' (Kytö and Rissanen, 1988). These are sets of two bytes, the first of which is a reserved character which indicates that the following one is not to be taken at its face value but as a special symbol which cannot be represented using the IBM extended ASCII character set to be found by default on all personal computers; in fact the Helsinki Corpus gets by with characters from the lower area of this set. So, for instance, the 'eth' character of Old English (a crossed 'd' which along with thorn, a Runic character, was used to represent the inter-dental fricatives of this stage of the language) is encoded as '+d'. Thorn itself is indicated as '+t', the medieval form of *g* 'yogh' is encoded as '+g', etc. The advantage of such a coding scheme is that of portability: texts can be transferred effortlessly from one environment to another, e.g. from a personal computer to a mainframe or a work station, from one operating system to another without entailing loss of data. The disadvantage should be obvious: one cannot see Old and Middle English symbols as they would be represented in printed form. For the linguist involved in analysis of medieval texts this is untenable in the long term. To alleviate the situation, a programme has been included in the *Lexa* set which will convert all escape sequences used for the Helsinki Corpus into single characters. If one then uses the special Old English character set supplied with the *Lexa* suite, then one actually sees Old and Middle special symbols as they appear in the printed forms of medieval texts. Furthermore, one can reverse the conversion of Helsinki texts, thus allowing portability to another environment at any time. A special keyboard driver and a printer driver for *WordPerfect*, as well as both dot matrix and

laser printer fonts are supplied with *Lexa* which allow one to enter from the keyboard, view on the screen and output on paper the Old and Middle symbols required for the earlier texts of the Helsinki Corpus.

4.3. A word on utilities

The third volume of the *Lexa* suite, which is concerned with general file management, bears the title *Utility library*. It embraces a series of programmes which perform various housekeeping tasks necessary for efficient data management on a personal computer. As such, the programmes are not primarily involved in corpus processing, but should nonetheless not be neglected by users. Mention should just be made here of the fact that CD-ROM drives are supported by the utility software, which means that the ICAME CD-ROM of English Language Corpora, which is available form the Norwegian Computing Centre for the Humanities, can be managed directly by the *Lexa* software.

Notes

1. General introductions to this field are provided by the collections by Aarts and Meijs eds. (1984 + 1986 + 1990) and Meijs ed. (1987). Particular discussions of software systems are to be found in Aijmer and Altenberg eds. (1991), the guide by Lancashire (1991) and the paper by Knut Hofland in Johansson and Stenström eds. (1991) as well as the collection by Kytö, Ihalainen and Rissanen eds. (1988). The application of corpus data to the question of linguistic variation is treated within a general framework in Biber (1988) and in the more specific context of computer corpora in Oostdijk (1988). For a useful bibliography on English computer corpora, see that by Altenberg in Johansson and Stenström eds. (1991).
2. The documentation is available as follows: Vol. 1: *Lexical analysis*, Vol. 2: *Database and corpus management*, Vol. 3: *Utility library*. Accompanying this are 4 microfloppies containing the software and sample data. In addition, both the documentation and the software will be available on the mainframe of the Norwegian Computing Centre for the Humanities so that interested parties can download both to their local computer by employing a commonly used file transfer protocol. Software updates and additions can be obtained in this manner.
3. The reason for the relatively large number of executable files is to allow users to select only those programmes for an area of corpus

processing which interests them and of course not to dismay the uninitiated by presenting them with very large programmes characterized by feature cluttering. Unified desktops for all the major programmes will hopefully drastically reduce the time required to acquaint oneself with the set. Many programmes interrelate, particularly with regard to data input and output. This is achieved automatically and should not disturb users.

4. *Lexa* can be employed gainfully with pre-tagged corpora (e.g. the LOB Corpus, see Johansson, *et. al* 1986), particularly for information retrieval tasks.
5. This format is catered for by much software which is intended for corpus data processing, e.g. the Oxford Concordance Program (Hockey and Marriott, 1980).
6. For those wishing to inquire further about tagging of corpus texts the following references might be useful. Francis (1980) offers a general discussion; Garside and Leech (1982) and Leech, Garside and Atwell (1983) discuss tagging of the LOB Corpus, Garside (1987) gives an introduction to the so-called CLAWS system while Akkerman, Meijs and Voogt-van Zutphen (1987) explains the methods used for the ASCOT project; Svartvik (1987) presents suggestions and a partial reappraisal of tagging proposals.
7. Prospective users are advised to acquaint themselves with statistics in general, especially with the types of statistical operations which are sensible in linguistics, see Butler (1985) for instance.
8. In 'Quantifying syntactic deviation in Synge's dramas', paper presented at the Workshop on Corpus Linguistics, Department of English, University of Innsbruck, January 1993.
9. See Hickey (1993b) for details of such structures in Irish English.

References

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

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

Aarts, Jan and Willem Meijs (eds). 1990. *Theory and practice in corpus linguistics*. Amsterdam: Rodopi.

Aijmer, Karin and Bengt Altenberg (eds). 1991. *English corpus linguistics: Studies in honour of Jan Svartvik*. London: Longman.

Akkerman, Eric, Willem Meijs, and Hetty Voogt-van Zutphen. 1987. Grammatical tagging in ASCOT. In *Corpus linguistics and beyond: Proceedings of the Seventh International Conference on English Language Research on Computerized Corpora*, ed. by Willem Meijs. 181-193. Amsterdam: Rodopi.

Altenberg, Bengt. 1991. A bibliography of publications relating to English computer corpora. In *English computer corpora: Selected papers and research guide*, ed. by Stig Johansson and Anna-Brita Stenström. 355-396. Berlin: Mouton de Gruyter.

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

Butler, Charles. 1985. *Statistics in linguistics*. Oxford: Blackwell.

Francis, W. Nelson. 1980. A tagged corpus – problems and prospects. In *Studies in English linguistics – for Randolph Quirk*, ed. by Sidney Greenbaum *et al.* 192-209. London: Longman.

Garside, Roger. 1987. The CLAWS word-tagging system. In *The computational analysis of English: A corpus-based approach*, ed. by Roger Garside *et al.* 30-41. London: Longman.

Garside, Roger and Geoffrey Leech. 1982. Grammatical tagging of the LOB Corpus: General survey. In *Computer corpora in English language research*, ed. by Stig Johansson. 110-117. Bergen: Norwegian Computing Centre for the Humanities.

Hickey, Raymond. 1993a. *Lexa. Corpus processing software*. 3 vols. Vol. 1: *Lexical analysis*. Vol. 2: *Database and corpus management*. Vol. 3: *Utility library*. Bergen: Norwegian Computing Centre for the Humanities.

Hickey, Raymond. 1993b. An assessment of language contact in the development of Irish English. In *Language contact and linguistic change*, ed. by Jacek Fisiak. Berlin: Mouton de Gruyter.

Hockey, Susan and Ian Marriott. 1980. *Oxford Concordance Program: Users' manual*. Oxford: Oxford University Computing Service.

Johansson, Stig, Eric Atwell, Roger Garside, and Geoffrey Leech. 1986. *The tagged LOB Corpus: User's manual*. Bergen: Norwegian Computing Centre for the Humanities.

Johansson, Stig and Anna-Brita Stenström (eds). 1991. *English computer*

corpora: Selected papers and research guide. Berlin: Mouton de Gruyter.

Kytö, Merja (compiler). 1991. *Manual to the diachronic part of the Helsinki Corpus of English texts.* Helsinki: Department of English.

Kytö, Merja, Ossi Ihalainen, and Matti Rissanen (eds.). 1988. *Corpus linguistics hard and soft.* Amsterdam: Rodopi.

Kytö, Merja and Matti Rissanen. 1988. The Helsinki Corpus of English Texts: Classifying and coding the diachronic part. In *Corpus linguistics*, ed. by Merja Kytö *et al.* 169-180. Amsterdam: Rodopi.

Kytö, Merja and Matti Rissanen. 1992. A language in transition: The Helsinki Corpus of English texts. *ICAME Journal* 16: 7-27.

Lancashire, Ian. 1991. *The humanities computing yearbook 1989-90: A comprehensive guide to software and other resources.* Oxford: Clarendon Press.

Leech, Geoffrey, Roger Garside and Eric Atwell. 1983. The automatic grammatical tagging of the LOB Corpus. *ICAME Journal* 7: 13-33.

Meijs, Willem (ed.). 1987. *Corpus linguistics and beyond: Proceedings of the Seventh International Conference on English Language Research on Computerized Corpora.* Amsterdam: Rodopi.

Oostdijk, Nelleke. 1988. A corpus linguistic approach to linguistic variation. *Literary and Linguistic Computing* 3/1: 12-25.

Svartvik, Jan. 1987. Taking a new look at word class tags. In *Corpus linguistics and beyond: Proceedings of the Seventh International Conference on English Language Research on Computerized Corpora*, ed. by Willem Meijs. 33-43. Amsterdam: Rodopi.

Some reflections on the question of teaching, from a corpus linguistics perspective[1]

Steve Fligelstone
University of Lancaster

1. What do corpora have to do with teaching?

In the past couple of years there is evidence of a huge leap in the perceived relevance of language corpora. This impression is supported by the growth in volume of publications (see Johansson, 1991:312), by the pattern of funding in recent years, including the emergence of large scale projects in Britain and the USA, by the fact that direct reference to corpora is made on the covers of several recent major dictionaries (that is to say, corpora have already made it onto the high street!), and by an apparent shift in attitudes among many who voiced scepticism about the value of corpora, to a point where few are now willing to dismiss their value and many seek to become better acquainted with techniques that may be applied to them. This is not really news to readers of the *ICAME Journal*, particularly as many of them are directly responsible for bringing about this desirable state of affairs!

However, being responsible for a change of attitudes and the popularisation of particular ideas does not guarantee one's satisfaction with their subsequent progression in the wider domain. Indeed there is often a contrary dynamic which causes the innovator to feel that his or her principles have been misunderstood and diluted as they have found favour with a wider audience. Corpora, it might be said, are currently 'in fashion', but fashions are inherently transitory. The real aim of corpus linguists has (I hope) been not so much to create a movement or a label by which to be known, but, rather, to reintroduce the principles and practice of empiricism into a field of enquiry which had been in danger of pursuing its flight of rationalist exploration past the point of

no return, and to promote the development of resources and practices which would greatly enhance the opportunities for empirical research. This rapid movement from the margins toward the mainstream of the ideas promulgated by corpus linguists for the past 30 years or so is the main reason why I suggest that we must focus carefully and imaginatively on the whole enterprise of teaching – whether of languages or of linguistics or natural language processing, since teaching is one of the major mechanisms for transmission of intellectual ideas, and it is the appropriateness and the effectiveness of what is taught, as well as *how*, and *to whom*, which ultimately will help to determine the success or otherwise of this reinvigoration of the empirical approach to language study.

Another reason for thinking about corpus linguistics and teaching in conjunction with each other is of course that corpora themselves, or rather, their exploitation, can actually assist in the teaching process – something already widely recognised but by no means an exhaustively explored area. The aim of this paper is to focus on some aspects of the question of corpora in teaching, and to encourage discussion of such issues amongst corpus linguists.

It may be useful to adopt a simple framework within which to try and assess the factors relevant to good teaching practice and to the development of the right sort of activities in the interests of linguistics and language study as a whole, from the perspective of one having a special interest in language corpora. I suggest that corpus-related activities can be loosely grouped into three categories which in practice interact in some quite interesting ways:

TEACHING ABOUT (i.e. teaching about corpora/corpus linguistics)
TEACHING TO EXPLOIT (i.e. teaching students to exploit corpus data)
EXPLOITING TO TEACH (i.e. exploiting corpus resources in order to teach)

Below I discuss each of these areas in turn.

'Teaching about' implies on the face of it a fairly traditional perspective in which, however, 'corpus linguistics' is identified as a field of study – a sub-discipline perhaps – and students are taught about it – the history, the theory, the objectives and the objections etc. That there is a basis for talking about 'Corpus Linguistics' is evidenced (for example) by the occurrence each year of the ICAME Conference, and by the frequent use of the term. But for reasons which hardly require elucidation, it is not necessarily desirable to be too rigid about the demarcation of

the field – in other words the 'Corpus Linguist' frequently is no such thing – he or she is a linguist who uses corpora as one of his or her tools (Johansson 1991:313). And likewise the distinction between Corpus Linguistics and 'Non-Corpus Linguistics' is not one we would wish to set in stone since many of the tenets of the former are (from an empirical perspective) tenets of language study *per se*. There appears to be a divergence of opinion even amongst those who use corpora most. In the books section of the last edition of *ICAME Journal*, one reviewer states: 'corpus linguists should spend less time talking to corpus linguists, and more time talking to other researchers whose work could be advanced by using corpora' (Sampson 1992:82), whilst another talks of 'the principles of the sub-discipline' and concludes that the book he is reviewing 'succeeds in showing that corpus linguistics "has developed into a sub-discipline in its own right"' (Schmied, 1992:101). These two statements do not actually contradict one another, and both accept the reality of the idea of a 'corpus linguist'. However, they do perhaps reflect a tension between different attitudes on the question of how the 'corpus community' should view and conduct itself. Ultimately, both tendencies should be reconcilable, for it is surely desirable that corpora are used both appropriately and widely. The former interest may well be best safeguarded by a community with a more or less well-defined 'special interest', but the latter certainly implies an outward looking approach. The question of the dissemination of 'Corpus Linguistics' can thus be seen to be partly a question of appropriate presentation and contextualisation. This has a bearing on one of the two recent Lancaster teaching endeavours referred to below. This is not, however, to dismiss the idea that Corpus Linguistics can or should be taught as a field in its own right, but neither should the idea be taken for granted.

Quite what Corpus Linguistics is, and how it fits into Linguistics as a whole, is something which will be discussed for a long time to come, but what there can be little doubt about is that when we reflect upon the best way to teach about Corpus Linguistics, we are in fact asking ourselves, 'how should we talk to the future research community?' It is thus interesting to note that consideration of the question of teaching forces us to consider a question right at the heart of our research activity. It is also important that we do not allow the relative newness of 'Corpus Linguistics' to cause it to be overlooked in the planning of the curriculum.

But if there is any truth in the old adage 'actions speak louder than words', then we should consider the possibility that what will have the most profound influence on our students is not what we say about

corpora, but what we actually do with them. From this perspective, the exploitation of corpora in the classroom, and the means we devise for students to gain 'hands-on' experience of corpora, are by no means incidental, but perhaps rather central to the future of language corpora as a widely used resource.

'Teaching to Exploit' is, on the face of it, a simple enough concept – theoretical and, more especially, practical training in the manipulation and uses of corpus data – but (precisely because of this practical emphasis) it is arguably the most important area in which to adopt the right teaching plans and practices. It is probably the area which has the greatest *psychological* impact on the student, and thus the potential to determine whether or not a student will embrace the techniques of corpus study and explore its scope, or whether they will simply pay it lip service and 'leave that kind of work to others'.

This is a problem largely related to the need to use computers – a situation with all too many well-known problems for which there are all too few well-understood remedies. A great deal of thought thus needs to go into the planning of activities which will encourage the learner to have the confidence and the motivation to come to grips with problems of potentially great complexity. There seem to exist different levels of mental involvement with computer technology, which do not necessarily combine in obvious ways. Some people will acquire considerable expertise with a complex desktop publishing or graphics package, and yet never compile a binary code in their lives. Others will instinctively feel that there is a sort of barrier which separates the computationally sophisticated world from the computationally naive – and those on one side can word-process and nothing else, whilst those on the other side can do just about anything. Yet others can handle the conceptual side of computing with relative ease, but never actually learn to do it. Such customs are, I think, deeply rooted and can become more, rather than less, ensconced over time. In trying to unravel this conundrum, one can do worse than ask the question: 'how did the person in the second case, above, actually learn to word-process in the first place?'. The likelihood is that, in many cases, they never actually thought of it as 'computing'.[2] If our attempts to involve others are to be as inclusive as possible, then we need to devote considerable effort to enabling those who 'cannot compute' to work with corpora. Fortunately, I think, the possibilities for achieving this have never been greater, due to advances in software design, but it is still a problem, and not only where students are concerned.

The question of access to appropriate resources is a key one. Many research establishments (until recently one could count our own amongst them), whilst supporting various levels of corpus-based research by their full-time staff, are poorly resourced for allowing ready access to corpora by the majority of staff or students for research purposes, let alone for teaching or open-learning purposes. The very reasonable question 'where can I find the corpora?' must all too often be greeted by a complex response which will make the enquirer wish they had never asked! In our own case that would have entailed instructing the enquirer to apply for a mainframe account, get acquainted with UNIX, some rather arcane programs, and central printing facilities. Without the resources to train people in these tasks, many will fail to engage in real corpus interaction. By stages we have been able to move much closer to a situation where we can give the hoped-for response: 'go to any of the labs, hit the icon which says "Corpus" and follow the instructions on the screen'. But such facilities require a degree of planning, which in turn entails some notion of how one envisages corpora being used.

Finally **'Exploiting to Teach'**. This is certainly the most explored of the three areas I have outlined, but I suggest it is also the least *exhaustively* explored. There is no contradiction: whereas the question of how to teach theory and how to teach practice are in principle questions of approach, and in a sense quantifiable problems, the question of how to exploit corpora for teaching purposes is certainly open-ended. It corresponds in fact to the question of how to exploit corpora for the purposes of *learning* about the language – we simply cannot predict at what point the question will cease to yield new and exciting answers. In any case, as Knowles (1990) points out, the distinction between teaching and study and research becomes very blurred in the case of corpus-based work, since even at the level of the most obvious kind of corpus use, one is entering into the realms of discovery, possibly novel discovery.

Perhaps it is useful to break down this area still further. On the one hand, what has been obvious to many researchers for some time is that much benefit can be derived in language teaching from the use of corpora as a means of determining what to teach. Many researchers, – for example Mindt (1988), Renouf and Sinclair (1988) – have shown the value of corpus data to inform teaching practice. Much work of this kind is reviewed in a valuable survey by Kennedy (1992). A key factor which emerges (and this is true of corpus linguistics generally)

is the way in which findings from corpus-based research contradict commonly held assumptions about language use.

But there is another, more direct, sense in which one can speak of corpus exploitation. Whether it is a case of 'concordances in the classroom' (see Tribble and Jones, 1990) or some other means of student-corpus interaction, there seems to be a growing concern with how corpora – a form of authentic language data – can be used as part of the teaching activity – and not only in the domain of second language teaching. It is in this area that I perceive a greater need for open discussion of methods. Many people are doubtless pioneering novel and imaginative techniques for teaching involving the use of corpora, but classroom practice remains one of those subjects which researchers tend to regard as a matter for private rather than public discussion. With such a potentially important new dimension to the business of language and linguistics teaching, I feel that such reticence is misplaced.

The following activities can all be carried out with a common or garden concordancer and a small corpus of 'newsy' material:

- Exploring the nature of idioms and collocations: for example, a sorted concordance of the word *life* will reveal a number of expressions of a more or less fixed nature. It may be instructive for groups of students to compare these expressions with their equivalents in other languages and to consider why some expressions become 'idiomatic'. One could envisage an activity such as this also being used in a language learning context.
- Topic preparation: generating concordances of certain key content words can be a very useful way for students (particularly younger ones) to gather ideas about a subject in order to write something about it – a subject such as 'energy' or 'war' for example. Again, this may be particularly valuable as a second-language exercise.
- Rhetorical questions: carry out a search on question marks. How many of the preceding 'questions' actually expect an answer?
- Critical perspectives on prescriptive grammar: for example use a pattern such as '. And' to look for examples of the 'forbidden' sentence-initial conjunction. Comparison of different text types might be illuminating.

These are simple illustrative examples of the way in which relatively simple corpus use by students can provide a variety of learning opportunities. They are conceived, of course, with less advanced students in mind. It does not matter that these issues have been tackled for years

by non-computerised means, and will continue to be. As with research, the corpus simply represents an *additional* tool that may be utilised by the teacher.

From a brief enquiry last year I learned of a number of activities currently being engaged in with University students (all references are to personal communications in 1992):

- Use of corpora as parser input data for undergraduate computer scientists (Atwell, Leeds University)
- Use of corpora to train 'non-computeers' in techniques of large data set analysis (Hearne, Western Washington University), and also in courses on Information Retrieval (Krovetz, University of Massachusetts)
- Use of syntactically marked corpus data in advanced pragmatics course (Ball, Georgetown University)
- Use of spoken corpus data in course on transcription techniques (Edwards, University of California, Berkeley).
- Student comparisons of corpora in undergraduate courses in language variation (Jappy, University of Perpignan)
- On-line student learning resource in Bulgarian language course (Hauge, University of Oslo)

These few examples (none of which, it may be noted, constitutes what might be termed a course in 'corpus linguistics') are doubtless only the tip of the iceberg, but it is at present impossible to assess how widespread the use of corpora currently is. The list does, however, demonstrate the breadth of uses to which corpora may be put.

One aspect of corpus use which I believe will receive much more attention in due course is the mode of interaction, and how this can be addressed by computational means, as distinct from human (training) resources. At present we are largely bounded, particularly those of us not engaged in Computer Science *per se*, by the possibilities offered by using existing software, but it is axiomatic of corpus-based research that the power of the computer has opened up previously unattainable research goals, and perhaps the same is true in the pedagogic arena. In other words, perhaps we should be starting to look beyond our familiar software and the potential for corpora to provide the underlying knowledge base for our courses, towards the eventual development of genuinely innovatory, computerised tutoring systems which facilitate a truly dynamic interaction between the learner and the data, in a way

that is geared towards the learner's needs. We can possibly expect to see considerable advances on this front over the next few years, the more so if we contribute to the effort.

2. *Recent experiences*

In spite of its relatively high profile as a research interest of the Linguistics Department, teaching of corpus linguistics at Lancaster has been concentrated very much, though not exclusively, at the postgraduate level. Indeed, it is probably fair to say that a significant proportion of the 'transmission' that has taken place has not been through conventional teaching at all, but through contact with the (inter-disciplinary) research unit UCREL, particularly its directors Geoffrey Leech and Roger Garside, who of course do lecture,[3] through a sort of 'wave effect' on less 'corpus-wise' academic colleagues and through the supervision of post-graduate research. Recently, however, we have found ourselves addressing questions of corpora and teaching more directly – in particular we have offered corpus-related tuition to first year undergraduates and to language teachers from outside the University. I shall now reflect upon these experiences and upon other developments which are currently unfolding.

One promising venture consisted of the inclusion of a corpus component in a two-week residential course for language teachers from various countries (see Gratze *et al* 1991). The three sessions offered included a brief, largely descriptive, introduction to language corpora, and two hands-on sessions using two very different kinds of concordancing software. The first hands-on session was in effect a tutorial in the use of the *Longman Mini Concordancer*. The reason behind this choice was that this piece of software works well enough and is simple enough to be taught more or less completely in a two-hour session and gave instant and untaxing access to some real corpus data.[4] In the second session less 'friendly' software was used to access larger and more richly encoded language corpora – by now a less unfamiliar object. An illuminating, if perfectly comprehensible, aspect of this teaching encounter was the ease with which this group of practitioners could grasp the potential usefulness of a corpus – once they had grown accustomed to their existence – this meaning that not too much time needed to be spent on the *teaching about* aspects of corpora – and it was possible to include such information in a quasi-anecdotal manner whilst carrying out practical tasks. Although the main activity of the course was in the *teaching to exploit* category, the desired outcome of the course would be a group of educators equipped to embark on an exploration of the

possibilities for *exploiting to teach*. However, things do come full circle here, since, returning to one of my opening remarks, one would hope for those embarked on such a course of action to have some awareness of theoretical issues concerning the data and methods they are using, and so the need to *teach about* arises after all.

In the other initiative we taught a group of 20 first-year undergraduates a non-trivial amount about corpus linguistics over the period of a 10-week option course which they took alongside a general introduction to linguistics. This, then, was primarily a *teaching about* activity, though a couple of hours of hands-on activity was included.[5] The novel thing about this series of sessions (as far as we were concerned) was that we did it under the auspices of a 'Language and Computers' (rather than 'Corpus Linguistics') course, and could thus lead the student on a path through the 'traditional' territory of computational linguistics to a point where some empirically based approaches to NLP and language research were looked at in some detail. There were some advantages to this – firstly, where recruitment was concerned, we were able to capitalise on a pre-established curiosity about 'Language and Computers' which could not be assumed to exist for a subject as obscure sounding as 'Corpus Linguistics'. Secondly, in pursuing the subject, we could approach questions of theory and rationale 'bottom up' so to speak, beginning with the problem rather than the solution, as we conducted our structured tour through a range of areas of computational linguistic research. Thirdly, and consequently, whilst due recognition was given to the relatively low profile accorded to corpus-based research during the past three decades, we did not feel it necessary to convey the idea that corpus linguistics was in any way *theoretically* marginal to computational linguistics, nor indeed that computational linguistics itself was a marginal activity.[6] This course was run for the first time in 1991-92 and was assessed as being a great success. In 1992-93 the course is being run again, with only minor modifications.[7]

In addition, the Lancaster University Linguistics Department (in association with the Computing Department) is now actively integrating corpus-related components into a number of other courses, at both undergraduate and postgraduate level. But there are other local symptoms of the growing impetus for teaching activities based around corpus methodology. Colleagues in the several language teaching departments are also engaged in serious if (hitherto) small-scale corpus-based research activities and are actively seeking out materials and methods to use with uninitiated students. More and more enquiries about corpus manipulation tools and techniques are being received. In response to this

we have recently held round table discussions leading to the formation of an informal grouping of local teaching practitioners with a common interest in corpora, though with widely differing experience in this area.

This initiative (which we call CAT – Corpora and Teaching) is still young, but tangible results have already ensued: several participants have already been encouraged to submit funding bids in this area – the first, part of a consortial bid for funding under the University Funding Council's TLTP[8] Initiative was unsuccessful, but a similar program has now been shortlisted for 'pilot' funding under a local higher education innovation scheme, along with a second corpus-based initiative devised by the Modern Languages Department. These pilot schemes, if they go ahead, will yield interesting insights. The Linguistics Department bid, in which the writer is involved, takes up the theme with which the previous section was concluded: the development of a genuinely and usefully 'corpus-driven' tutoring system.[9] Our chosen 'problem' is that of training in grammar, using largely open-learning methods, in order to redress the great (and increasing) divergence which is evident in the grammatical awareness of students attending University for the first time. Our approach, very roughly speaking, is to create an on-line resource which, as well as containing information of an expository nature, will present the student with graded tasks, such as particular levels of part-of-speech labelling or grammatical constituent recognition. The fact that the 'task-generator' will be using as its underlying resource annotated treebanks, will mean that it cannot only assess and categorise the student's progress, but precisely because of this categorisation, will be able to calculate a profile of the student's strong and weak areas of knowledge. This in turn will enable the tasks generated to be weighted towards the areas in which the student displays most difficulties. Eventually, we envisage not only the culling of relevant text fragments from a corpus, but the real-time processing of learner input. The methodology implies a model somewhat at variance with the text-book approach, as a pre-planned program of instruction is largely replaced by a student-centred program of exercises. The project in fact constitutes a first attempt to implement some of the ideas expounded several years ago by Geoffrey Leech (1986) in his paper on the educational applications of automatic grammatical analysis. We hope to pilot a system of this kind with the student intake in the coming academic year.

A second outcome of the CAT discussions has been a series-of tutorial workshops and demonstrations of various software packages ranging from retrieval programs through hypertext authoring systems to on-line conferencing systems (which we use to talk about corpora!) and qualitative

analysis packages. This latter initiative has been popular and seems likely to continue in some form. Its ultimate success will be demonstrated if, as a result, the number of people able to offer such training grows, enabling more people to benefit in due course. It is by such means, perhaps, that corpus linguistics, however defined, can 'break (even further) out of the closet'.

These sessions were rendered considerably more satisfying than they might otherwise have been, by the inauguration at about that time of a general Linguistics PC laboratory, in which various items of linguistic software were presented as icons in a Windows™ interface, with a range of corpora stored centrally (and safely) on a network server. As part of my ESRC-funded research (cf note 1) I am now engaged in the design and implementation of a corpus access 'front end', which we term a 'Corpus Workstation', whose aim is to integrate various packages within a single system with the addition of help facilities and other features not provided within the packages themselves. The purpose is to aid still further the practical problem of software familiarisation, and encourage the wider use of corpora in research, though it is to be hoped that the enhanced facilities will also be of benefit in a teaching context.

I started by saying that I felt teaching to be a key determinant of the overall long-term impact of the ideas which underlie and constitute the field of corpus linguistics. But in holding this view, I am also uncomfortably aware of how little I feel I know of what can and ought to be done to harness the enthusiasm of students, and of what others in the same predicament do. I have also formed the impression that I am not alone in this outlook. Drawing on the discussion in this paper so far, and on the kinds of ideas that have emerged in discussions on the subject, I would propose the following as an incomplete list of topics that may merit consideration by the 'ICAME community':

- How can we make corpora more relevant at undergraduate level (and how far should we do so)?
- Is there a case for teaching 'Corpus Linguistics' as a subject, and if so, when?
- How do we best bring about the wider dissemination of corpus manipulation skills (including amongst teachers!)?
- Are there particular problems concerning corpora and the computationally 'naive', and if so, how should they be tackled?
- Is there a need for a survey of corpus-related teaching activity?

Many readers of the *ICAME Journal* are regularly involved in teaching activities and in associated planning. I conclude by suggesting that such activity is, or at least should be, emerging as a new (additional) focal point for committed corpus linguists and that it would therefore be appropriate for this new focus to be reflected in contributions on such topics at future ICAME gatherings and in this Journal. It would be a shame if, as a research community with some good ideas, we did not take advantage of opportunities to discuss the best ways to pass those ideas on to others.

Notes

1. The writer is engaged on the ESRC project 'Lancaster Database of Linguistic Corpora' one aim of which is to research and develop resources designed to facilitate corpus-based research. This article is based on a discussion paper written for the 13th ICAME Conference, held in Nijmegen, 1992.
2. It is amazing how many people who answer 'no' to the question, *have you had any practical experience of computing*? subsequently reveal that they have frequently used computer programs such as word processors and/or games. Of course, there are those who are terrified even by this prospect.
3. In *Linguistics* and *Computer Science*, respectively.
4. Continued use of this program has convinced me that of all the available software it is the program best suited to 'cutting one's teeth' on techniques of concordancing. Its ease of use combined with its useful range of features make it highly effective as a vehicle by means of which to learn the basic concepts and techniques of corpus interrogation, though its limitations (particularly of text size) quickly lead one to require use of a more powerful program.
5. Not surprisingly, perhaps, but I think encouragingly, a significant proportion of the students felt that this had been too little.
6. Echoing a similar sentiment perhaps, the *ICAME Journal* has recently acquired a distinctly non-marginal sounding gloss on its front cover.
7. This course was devised and run by Gerry Knowles, Tony McEnery, Andrew Wilson and Steve Fligelstone.
8. Teaching and Learning through Technology Programme.
9. The other members of the team working on this proposal are Tony McEnery, Geoffrey Leech and Jenny Thomas.

References

Aijmer, Karin and Bengt Altenberg (eds). 1991. *English corpus linguistics: Studies in honour of Jan Svartvik*. London and New York: Longman.

Gratze, Charlotte, Franz Mittendorfer and Bernhard Kettemann. 1991. Computers in English language education and research. Report on British Council Specialist Course 9127, 3-15 April 1991, University of Lancaster. *ReCALL Bulletin* (Journal of the CTI Centre for Modern Languages, Hull University). No. 5. 15-16.

Johansson, Stig. 1991. Times change, and so do corpora. In Karin Aijmer and Bengt Altenberg (1991). 305-314.

Kennedy, Graeme. 1992. Preferred ways of putting things with implications for language teaching. In *Directions in corpus linguistics: Proceedings of Nobel Symposium 82, Stockholm, 4-8 August 1991*, ed. by J. Svartvik. 335-373. Berlin: Mouton de Gruyter.

Knowles, Gerry. 1990. The uses of spoken and written corpora in the teaching of language and linguistics. *Literary and Linguistic Computing*, Volume 5, No. 1. 45-48.

Leech, Geoffrey. 1986. Automatic grammatical analysis and its educational applications. In *Computers in English language teaching and research*, ed. by G. Leech and C.N. Candlin. 205-214. London: Longman.

Mindt, Dieter. (ed). 1988. *EDV in der angewandten Linguistik. Ziele – Methoden – Ergebnisse*. Frankfurt am Main: Diesterweg.

Sampson, Geoffrey. 1992. Review of K. Aijmer and B. Altenberg (1991). *ICAME Journal*, No. 16. 81-82.

Schmied, Josef. 1992. Review of N. Oostdijk. 1991. Corpus linguistics and the automatic analysis of English. *ICAME Journal*, No. 16. 101-103.

Sinclair, John and Antoinette Renouf. 1988. A lexical syllabus for language learning. In *Vocabulary and language teaching*, ed. by R. Carter and M. McCarthy. 140-160. London: Longman.

Tribble, Chris and Glyn Jones. 1990. *Concordances in the classroom: A resource book for teachers*. London: Longman.

Reviews

Jan Aarts and **Willem Meijs** (eds) *Theory and practice in corpus linguistics*. Amsterdam: Rodopi (Language and Computers: Studies in Practical Linguistics No. 4). 1990. iii + 254 pp. ISBN: 90-5183-174-9. Reviewed by **Graeme Kennedy**, Victoria University of Wellington.

Corpus linguistics covers a number of quite different activities including corpus making, the development of software to tag, parse or analyse corpora, linguistic descriptions based on corpora, and various applications such as the exploration of automatic natural language processing and language pedagogy. The eleven papers in this book range widely over a number of these activities.

Most of the papers are particularly concerned with the development of methodology in corpus research, rather than with theory or theories of language as such. Indeed it can be argued that apart from the emphasis on the probabilities of occurrence of linguistic items in texts one of the major contributions of corpus linguistics to linguistic science is in the methodologies used for arriving at descriptions of language. Although many of the papers include work in progress on aspects of linguistic description, there is less emphasis on the goals of corpus linguistics and on applications of corpus research. The papers are generally well-written, succinct and interesting, although there are more typos than might be expected. The editors have provided an excellent Preface which gives a brief overview of the contents of the book.

In his paper, Kaye describes software developed to build and analyse quite large (1 million word), plain, glossed, transcribed or tagged corpora on an IBM PC/AT. As with WordCruncher, the new software uses indexing which makes possible very rapid retrieval of data including concordancing, with economical use of disk space. The particular strengths of the new software will, of course, be revealed through use, which is facilitated in the best tradition of computer corpus linguistics by being made available to academics for non-commercial use at no charge. Belmore's paper suggests from a user's perspective just how far the hardware and software advances of the 1980s have set the stage for Macintosh-based research on corpora. As with several of the papers, the reader is left with the impression that the next challenge is to formulate research questions which fully exploit the opportunities now made available by the hardware and software.

The papers dealing more directly with linguistic description cover a

variety of topics. Campbell reports the development of a method by which syllable length in a 52,000-word corpus was analysed with a view to improving the prosodic naturalness of text-to-speech synthesis. In the absence of the availability of completely automatic (and possibly non-linguistic) analysis of spoken language, the study confirms the importance of painstaking manual analysis of corpora combined with the use of software, which it is claimed, can iteratively improve on the descriptive accuracy of the speech rate rules derived from the text.

The papers by Stenström and Janssen grapple with meaning. Stenström's paper outlines a sophisticated model of discourse signals in sentences, but is a salutary illustration of the difficulties facing researchers who seek to capture the range of functional complexity in discourse structure, involving as it does the multiple use of particular signals in different contexts. Whether automatic or manual tagging and analysis is used, the overriding problem remains of losing sight of the semantic forest among the grammatical and pragmatic trees.

Janssen, on the other hand, in spite of her title ('Automatic sense disambiguation'), has a somewhat narrow focus on meaning, describing work in progress on a procedure which uses the semantic information given in the computerized *Longman Dictionary of Contemporary English* to identify the particular sense of words in a corpus. As the author acknowledges however, automatic semantic analysis will need, among other things, a probabilistic element to complement the haphazardness of trial and error. Even if a way is found to achieve lexical disambiguation computationally, there is still the major further step of working out how to establish automatically the combinatory sense in propositions which is characteristic of natural language.

Souter's paper describing work on a corpus-based systemic-functional approach to natural language processing suggests that there is a possibility that 'a comprehensive grammar for English would be as open-ended as its vocabulary' (194). If indeed the number of syntactic rules needed to describe language in use is as vast as is suggested here, computational analysis involving probabilistic parsing and very fast processing provide the best hope for improved descriptions of English suitable for automatic natural language processing. The major theoretical issue of whether corpus-based computational grammars can replace competence or idealized grammars remains open. The aims of this project are important, but the author acknowledges that there remains much work to be done on such fundamental matters as finding ways of capturing discontinuity in trees, and the problem of ellipsis.

Briscoe's paper, on the other hand, paints a somewhat more optimistic

picture of grammatical regularity. He argues that there is greater regularity in NP types than other researchers have suggested. He claims that a relatively small number of rules (some of which are rarely used) can successfully parse about 97% of NPs in a 10,000-word sample of NPs from the LOB Corpus tree bank. It remains an open question, however, whether or not attempting to achieve an incremental improvement of grammatical rules through corpus analysis may be a doomed enterprise as Souter and others might suggest in the face of the evident complexity of natural language.

Two of the most interesting papers in the book are on what constitute the units of natural language. The existence of prefabricated routines or collocations which straddle lexis and grammar has long been recognized. Papers by Altenberg and Eeg-Olofsson and by Kjellmer describe projects designed to explore the nature and structure of these multi-word expressions. Altenberg and Eeg-Olofsson describe a project to study significant collocations in the London-Lund Corpus (LLC) using an IBM PC where possible. Although the LLC is unlikely to be big enough at half a million words to fully explore the nature of collocations in spoken English, the machine-readable format, careful prosodic transcription and varieties of speech make it a most suitable basis for the development of methodology as well as for the description of significant patterning.

The systematic design of this ambitious project, encompassing a series of stages which, among other things, eliminate 'phraseologically irrelevant combinations' (e.g. *of the*, *it a*), should be of considerable use for other researchers working with larger corpora including the forthcoming International Corpus of English. The authors have outlined a worthwhile and focused series of research questions to be addressed by the project, relating to linguistic theory, grammatical and lexical description, psycholinguistics, stylistics, computational linguistics and language learning and teaching.

Kjellmer has been a major contributor to corpus-based research on collocation and in this study of the Brown Corpus he explores the question of what lexical factors predispose towards collocability. He has found, for example, that adjectives and adverbs prove to be much less collocational than verbs or singular and mass nouns. In order to cope with the range of data, this project includes words which have only one tagged grammatical function and thus excludes collocations involving many high frequency items such as conjunctions, prepositions and determiners which have multiple functions. This paper shows the importance of automatic computational analysis to discover general tendencies in collocational structure. Studies of collocations which involve

computer-assisted manual analysis, on the other hand, can lose sight of such general tendencies by showing that prefabricated routines can be derived from almost all grammatical combinations.

The remaining papers focus on other aspects of the description of English. Ihalainen has studied dialectal speech from the Helsinki Corpus and his preliminary work confirms that computer corpus-based research makes possible the production of a much more accurate, detailed and insightful grammar of English dialects than has hitherto been possible. It is pleasing to note that the use of the CLAWS tagger on transcribed spoken dialectal texts proved to be satisfactory in this study. Ihalainen makes the interesting methodological claim, based on his research on relative clauses, that a corpus of about 40,000 words per speaker is big enough for systematic analysis of all but the rarest structures.

Wikberg used WordCruncher with a small corpus to study theme-rheme and lexical cohesion among particular words in photographic manuals. By following the occurrence of particular words through the corpus, he shows how theme and rheme interact. The finding that in these texts the theme contains much more information than expected invites re-consideration of the nature of theme and rheme.

Overall, then, the volume contains reports on a variety of topics and fields of activity within corpus linguistics. As hardware has become more available in the form of powerful personal computers, and software for tagging, parsing and analysis of text has become more accessible and user-friendly, the need for research agendas has become more striking. The papers in this volume demonstrate that a number of such agendas now exist and are being vigorously pursued over a wide range of areas of enquiry. In this way corpus linguistics contributes not as a separate branch of linguistics but with methodologies which contribute to the language sciences as a whole.

John Sinclair. *Corpus, concordance, collocation*. Oxford, New York etc.: Oxford University Press, 1991. 179 pp. ISBN 0-19-437144-1. Reviewed by **Kay Wikberg**, University of Oslo.

The Cobuild Project is well known to the readers of this journal as a major source of information on present-day English lexis and grammar. John Sinclair's new book is based on previously published papers, which have now been edited to sum up his interesting ideas about the study of the area between lexis and grammar. Multi-million-word corpora now give us access to the sort of lexico-grammatical study that Sinclair and

Halliday signalled their interest in many years ago. This book is very much about delicacy in descriptive linguistics, about exploring the borderland between grammar and lexis and making proper use of the evidence provided by all the data available.

Sinclair criticizes traditional linguistics for being too narrow and not abstract enough. He further claims that the bottom-up analysis of traditional linguistics fails to get to 'whole texts of any length and complexity, and where it does it seems unable to maintain connection between the large units and the small ones.' (p. 9) Certainly, text linguists have done some work in this area where traditional linguists have failed, which would have deserved a comment. I would also have liked to learn more about how Sinclair's emerging new patterns can be used to throw light on the connection between the microstructure and macrostructure of a text.

Drawing on the evidence of concordances largely based on the Birmingham Collection of English Texts, Sinclair sets out to give the term 'collocation' a new significance. He does so in a series of case studies in Chapters 3-6, and 8. Chapter 1, 'Corpus Creation,' is a guide to corpus compilers. Sinclair advises them to drop specialized material, to go in for 'many millions of words' (p. 19), and to include complete texts rather than text fragments. He is not the only one to have been frustrated with the bits and pieces found in many of the text samples in the standard corpora. Admittedly, neither text linguistic nor semantic research is very rewarding if the data are taken from the standard corpora only.

Sinclair is no longer impressed with what he calls the one-million word 'sample corpora'. It would have been interesting to know if that attitude applies to the forthcoming International Corpus of English as well, which is in part based on similar principles. Sinclair's criticism of the 'largely intuitive criteria' used to determine the genres of the sample corpora is justified. On the other hand, the availability of the tagged LOB Corpus has been a great asset to grammatical and vocabulary research, and its value will hardly diminish until progress in software allows the ordinary non-programmer to do his own tagging.

Chapter 2, 'Basic Text Processing,' deals with elementary concepts such as words, word-forms, types of frequency lists, and word frequency profiles. Chapter 3, 'The Evidence of Usage,' is about lexicography. The point Sinclair is making here is that concordances can now provide statistical evidence on how the separate word-forms and the senses of a given lemma are distributed. The new generation of corpora will allow lexicographers to order the senses according to frequency, to include

only such items as are in use, and will provide guidelines for which words and senses to scrap. My only objection to this is that even in the future the users of ordinary dictionaries may want to know the senses of words which were frequent one or two generations ago. In the light of all the new data coming from the 'monitor corpora' lexicographers will have to make difficult judgments as to what to include and what to drop.

Chapter 4, 'Sense and Structure in Lexis,' examines the hypothesis that

> there is a close correlation between the different senses of a word and the structures in which it occurs. 'Structures' includes lexical structure in terms of collocations and similar patterns. 'Senses of a word' includes the contribution that a word may make to a multi-word lexical item. (p. 53)

Sinclair examines the lemma *yield*, of which there are 125 instances in his corpus of 7.3 million words. To compare Sinclair's figures with those in the standard corpora I made the following table:

	BROWN	LOB	KOLHAPUR	LONDON-LUND
yield	35	42	72	10
(N)	(18)	(16)	(54)	(6)
(V)	(17)	(26)	(18)	(4)
yielded	12	6	15	–
yielding	8	4	12	–
yields	7	10	16	–
(N)	(3)	(3)	(15)	
(V)	(4)	(7)	(1)	
TOTALS	62	62	115	10

(The London-Lund Corpus figures have been doubled to match the other corpora.) The table contains several surprises. One is the high frequency of *yield* in all the written corpora, altogether 239 instances in 3 million words. The expected frequency in each corpus calculated on the basis of the Birmingham Corpus is 17. Part of the explanation is to be found in the skewed distribution of *yield*, i.e. all the written corpora contain some texts with many occurrences of *yield*. Thus in the Brown Corpus B21 deals with *yield(s)* of x megaton, LOB J19 contains the collocations *yield conclusions/results*, and KOLHAPUR E36-E37 are about wheat production. Another explanation might be that *yield* is a word which is becoming less frequent. It is certainly rare in spoken academic discourse,

but then we all know that the speakers performing in the London-Lund Corpus do not represent a Department of Agriculture.

Chapter 5, 'Words and Phrases,' is an analysis of word combinations made up of *set* + particles. As regards *set in*, Sinclair observes that

> The most striking feature of this phrasal verb is the nature of the subjects. In general, they refer to the unpleasant states of affairs. (p. 74)

This is the kind of interesting information that very large corpora will help us to bring to light.

Of is a word hardly anybody would look up in the dictionary, but Sinclair devotes 17 pages to it in Chapter 6, 'The Meaning of Lexis and Grammar.' He argues that his data support an analysis of *of* as a 'one-member class ... where grammar and lexis join' (p. 83). He analyses N_1+*of*+N_2 nominals in some detail, looking particularly at types of N_1. Again it is the access to very large corpora that enables Sinclair to find new form-sense relations.

Chapter 7, 'Evaluating Instances,' calls into question the separation of lexis and syntax in grammar. Sinclair would rather

> widen the domain of syntax to include lexical structure as well, and call the broader domain *structure*. (p. 104)

This stands out as his most explicit statement on how he would like to extend traditional linguistics. Since sense and structure are inseparable, lexicographers need typical citation forms combining sense and structure. The starting point for the search is concordances. Sinclair concludes that

> Most everyday words do not have an independent meaning, or meanings, but are components of a rich repertoire of multi-word patterns that make up text. This is totally obscured by the procedures of conventional grammar. (p. 108)

Chapter 8, 'Collocation,' further develops the analysis of word co-occurrence and is important because this is where Sinclair describes his two models of interpretation of vocabulary, i.e. the 'open choice' and the 'idiom principle', which contrast sharply with each other. It is obviously the latter Sinclair goes in for:

> The principle of idiom is that a language user has available to him or her a large number of semi-preconstructed phrases that

constitute single choices, even though they might appear to be analysable into segments. (p. 110)

The idiom principle has been with us for some time already (cf Pawley & Syder 1983), although not expressed in this way, but what is new is 'the progressive delexicalization' of high-frequency words, and, consequently, that 'normal text is largely delexicalized, and appears to be formed by exercise of the idiom principle, with occasional switching to the open-choice principle.' (p. 113)

The last section, 'Words about Words,' deals with the structure of explanations in the Cobuild dictionary, and therefore deviates somewhat from the main themes of the book. It can be seen as a defence of the Cobuild explanatory style, which makes use of ordinary language, and which therefore, according to Sinclair, is open to entailments, implications and inferences like any other type of discourse. The book ends with a useful glossary of elementary terms.

A book like Sinclair's has been needed for some time. Although it would undoubtedly have been more coherent and less repetitive if it had been written from scratch, the ideas are fresh and stimulating. A practical problem with the monitor corpora, which Sinclair is a proponent for, is that rather few centres for corpus research will have the resources to set up such gigantic projects. One would therefore hope that some of the information will flow on to researchers who have to make do with less, such as via communication networks or CD-ROM disks and with software that allows you to search tens of millions of words of discourse at a time. Only then can we start evaluating the implications of Sinclair's statements to the full.

Reference

Pawley, Andrew & Frances Hodgetts Syder. 1983. Two puzzles for linguistic theory: Nativelike selection and nativelike fluency. In Jack C. Richards & Richard W. Schmidt (eds.), *Language and communication*. London & New York: London. 191-226.

Shorter notices

Designing a corpus of Cameroonian English

David Tiomajou
University of Yaounde, Cameroon

1. Introduction

As an imported language in many countries of the world, English has undergone local changes that have evolved into distinct and different varieties of the language. These varieties call for linguistic and socio-linguistic studies of their particular features. It is in this connection that, in February 1992, a research link was established between the University of Yaounde in Cameroon and the University of Birmingham in the U.K. The chief objective of such a link is to encourage linguistic research and enable Cameroonian English lecturers to have some academic support from their British counterparts at Birmingham University. A research project was accordingly initiated in February 1992 at Yaounde University, under the local supervision of the Head of the Department of English and the distant support of the University of Birmingham and the British Council.

2. Objectives of the project

The main objective of the project is to set up a million word corpus of Cameroonian main stream 'educated English', entirely based on written usage. Renouf (1986) asserts that

> A collection of texts, or corpus, can be processed by computer to produce information both statistical and linguistic, which is of use to the language teacher, the material writer, the lexicographer and the linguistic researcher...
>
> (Renouf, 1986:1)

and it is hoped that a corpus of Cameroonian English will be useful in a whole range of ways, including the following:

- It will provide a textual basis for a quantitative study of Cameroonian English.
- It will offer a database for the description of the main features and problems inherent in the variety of English which is spoken in Cameroon.
- It can be used as a source of authentic material for TEFL and TESL in Cameroon.
- It will be a database for comparative studies of Cameroonian English with other varieties of English like Nigerian, Indian, British or American English.

3. *Background*

As Cook (1991) rightly observes

> To most people in England it seems remarkable that someone can use more than one language in their everyday life; to most people in the Cameroon there is nothing surprising in using four or five languages in the course of a day.
>
> (Cook, 1991:102)

Cameroon is in fact a complex multilingual country in Sub-Saharan Africa, with 239 languages (see *Atlas Linguistique du Cameroun*, 1983), divided as follows:

- 236 local languages,
- 2 official languages (English and French),
- a sociolinguistically powerful Lingua Franca (Cameroon Pidgin English).

With no official status, the 236 local languages are mostly used in families and traditional settings. Although English and French enjoy an official status in the country, the national policy of French-English bilingualism instituted since 1961 has never had a nationwide systematic planning with clear-cut objectives. Moreover, Cameroon Pidgin English (CPE) remains by far the most widely spoken language throughout the country. Such is the background against which the English language is acquired, learnt and taught, both as a foreign and a second language, in Cameroon. It should be noted that for some Cameroonians, English may be the third or the fourth language.

4. *Design parameters*

Renouf (1987) observes that

> When constructing a text corpus, one seeks to make a selection of data which is in some sense representative, providing an authoritative body of linguistic evidence which can support generalisations and against which hypotheses can be tested.
>
> (Renouf, 1987:2)

The term 'representative' is controversial, since it is impossible to know the totality of a language, and so impossible to produce a microcosm of it. However, a corpus may be said to be 'representative' in so far as it attempts to reflect a wide range of the salient text types and linguistic choices found in the language in question.

Accordingly we have adopted the following design parameters for the corpus of Cameroonian English:

- The corpus will cover a broad range of the local written usage of Cameroonian English.
- It will include fiction and non-fiction; and popular, scholarly and literary texts.
- The texts will be selected from general domains.
- The texts will primarily be written by Anglophone Cameroonians.
- The texts will be from 1990 and beyond.
- The data will include texts from female as well as male writers.
- The text length will be between 2000 and 5000 words.

5. *Text categories*

At the beginning of the project, it was felt that the corpus of Cameroonian English should be as original as possible, not copying any existing corpus models but reflecting the particular sociolinguistic, cultural and educational realities of the country. However, after investigating the field of corpus linguistics in more detail, we gradually came to the conclusion that we would probably benefit from a more conventional format for the corpus. We therefore decided to include the following text categories, which are broadly inspired by the text categories in the Brown and LOB corpora and the corpus project of Australian English:

A. Official press
B. Private press
C. Novels and short stories
D. Religion
E. Tourism
F. Official letters
G. Personal letters
H. Students' essays
I. Government documents and memos
J. Advertisements
K. Miscellaneous

6. Constraints

6.1. Personnel

The first difficulty for the project is the lack of personnel dedicated and trained for the task. At the moment, only three people are actively involved in the project: the Head of the Department of English who is the coordinator of the project; and two part-time corpus builders who also do the keyboarding of the data.

6.2. Resources

With no internal maintenance budget, the project relies heavily on the support of the British Council and the University of Birmingham. Cameroon lacks basic computing equipment such as floppy disks, which have to be ordered from Britain. In addition there is very limited availability of electronic text anywhere in Cameroon and no local facilities for text scanning. This means that the texts are having to or will have to be keyboarded, with all the difficulties that this process entails, especially for people with no real clerical expertise.

6.3. Data

Given that the corpus will include only written material, it will not be easy to acquire the data needed because Cameroonian society, like many of its African counterparts, has an orally-based culture, and an acute shortage of reading resources, libraries, printing and mass communication facilities.

6.4. Selection

Another difficulty that we can foresee with regard to the project is that of text selection. The notion of 'educated English' is a complex and rather open-ended one. In addition, the dualism of the Anglophone/Francophone society, which leads to the English language having a status of both second and foreign language in the country, may cause problems in the sense that some Francophone Cameroonians are very proficient in English, and vice versa.

7. Prospects

It is hoped that the training received by one Cameroonian in corpus linguistics and text processing at the University of Birmingham in Summer 1992 will benefit the project, and that progress will accelerate. However, the local team will continue to rely on its British counterpart.

It was initially planned that the project would consist of a 4-stage activity:

Stage 1: training of staff, research contacts, collection and selection of data;

Stage 2: building of one quarter of the corpus;

Stage 3: building of one half of the corpus;

Stage 4: building of the final quarter of the corpus.

It was also estimated that the project would run from March 1992 to December 1993. Having assessed the technical problems and, the computational and linguistic skills involved, however, it now seems that the project may last longer than the research team originally estimated, probably extending into 1994. So far about three hundred thousand words of the various text types have been keyed in.

8. Conclusion

In identifying the importance of the International Corpus of English, Greenbaum (1991) noted that

> The project will undoubtedly provide valuable information on the use of English in many countries, in most of which there have never been systematic studies, and it will provide the basis for international comparisons. It will stimulate insights into the socio-

linguistics of English nationally and internationally, and offer data for sociolinguistic theory.

(Greenbaum 1991:91)

Though not as ambitious as the ICE project, the Cameroonian English corpus project is expected to provide a reliable source of data that will motivate and foster linguistic and sociolinguistic studies in a variety of English with reference to a very multilingual context, namely that of the Republic of Cameroon.

Note

I would like to express special gratitude to Antoinette Renouf for her advice and suggestions.

References

Cook, Vivian 1991. *Second language learning*. London: Edward Arnold.

Greenbaum, Sidney 1991. The development of the International Corpus of English. In *English corpus linguistics*, ed. by Karin Aijmer and Bengt Altenberg. 83-91. London: Longman.

Green, Elizabeth and Pam Peters. 1991. The Australian Corpus Project and Australian English, *ICAME Journal* 15, 37-53.

Renouf, Antoinette 1986. The exploration of a computerised corpus of English text. Paper presented at the VIIIème Colloque du G.E.R.A.S.

Renouf, Antoinette 1987. Corpus development. In *Looking up: An account of the COBUILD project in lexical computing*, ed. by John M. Sinclair. 1-40. London & Glasgow: Collins ELT.

Société Internationale de Linguistique. 1983. *Atlas Linguistique du Cameroun*. Yaounde.

The SUSANNE Corpus

Geoffrey Sampson
University of Sussex

Colleagues needing the use of a grammatically-analysed corpus of English may like to know that Release 1 of the SUSANNE Corpus is now complete, and is freely available from the Oxford Text Archive via anonymous ftp to any machine connected to the Internet. Instructions for retrieving a copy of the Corpus are given at the end of this announcement.

The SUSANNE Corpus has been created, with the sponsorship of the Economic and Social Research Council (UK), as part of the process of developing a comprehensive NLP-oriented taxonomy and annotation scheme for the (logical and surface) grammar of English. The SUSANNE scheme attempts to provide a method of representing all aspects of English grammar which are sufficiently definite to be susceptible of formal annotation, with the categories and boundaries between categories specified in sufficient detail that, ideally, two analysts independently annotating the same text and referring to the same scheme must produce the same structural analysis. The SUSANNE scheme may be likened to a 'Linnaean taxonomy' of the grammatical domain: its aim (comparable to that of Linnaeus's eighteenth-century taxonomy for the domain of botany) is not to identify categories which are theoretically optimal or which necessarily reflect the psychological organization of speakers' linguistic competence, but simply to offer a scheme of categories and ways of applying them that make it practical for NLP researchers to register everything that occurs in real-life usage systematically and unambiguously, and for researchers at different sites to exchange empirical grammatical data without misunderstandings over local uses of analytic terminology.

The SUSANNE Corpus comprises an approximately 128,000-word subset of the Brown Corpus of American English, annotated in accordance with the SUSANNE scheme. The SUSANNE analytic scheme is defined in detail in a book by myself, *English for the Computer*, forthcoming from Oxford University Press, and briefly in a documentation file which accompanies the Corpus. The Chairman of the Analysis and Interpretation Working Group of the US/EC-sponsored Text Encoding Initiative has proposed the adoption of the scheme as a recognized TEI standard. The

SUSANNE scheme aims to specify annotation norms for the modern English language; it does not cover other languages, although it is hoped that the general principles of the SUSANNE scheme may prove helpful in developing comparable taxonomies for these.

Regrettably, Release 1 of the SUSANNE Corpus is not a 'TEI-conformant' resource, though aspects of the annotation scheme have been decided in such a way as to facilitate a move to TEI conformance in later releases. The working timetable of the Initiative meant that relevant aspects of the TEI Guidelines were not yet complete at the point when the SUSANNE Corpus was ready for initial release; delaying this release would have been unfortunate.

Although the SUSANNE analytic scheme is by now rather tightly defined, Release 1 of the SUSANNE Corpus undoubtedly still contains errors despite considerable proof-checking. It is intended to correct these in later releases; I should be extremely grateful if users discovering errors would notify me, preferably by post rather than e-mail.

The SUSANNE Corpus consists of 64 data files (each comprising an annotated version of one Brown text), together with a documentation file. However, the versions held by the Oxford Text Archive are compressed, in order to reduce file transfer time, into single files in two alternative formats, suitable for Unix users and for users who have access only to a PC. The procedure for retrieving a copy of the Corpus in either case is as follows:

From a machine on the Internet, type either:

ftp black.ox.ac.uk

or, since the Archive is not yet in many official name tables:

ftp 129.67.1.165

When connected, you will be prompted for an account name, to which you should respond:

ftp

or:

anonymous

You will be asked to supply a password, in response to which you should type your e-mail address. After this is accepted, your first command should be to move to the directory containing the Text Archive files, by typing:

cd ota

To see a list of the files and directories currently available, type:

```
ls
```

All files relating to the SUSANNE Corpus are kept in the directory 'susanne', so your next command should be:

```
cd susanne
```

Apart from a README file containing the instructions which you are currently reading, this directory contains the two alternative compressed versons of the SUSANNE Corpus. To retrieve a copy of the corpus, if you are a Unix user, type:

```
get susanne.tar.Z
```

Having successfully transferred a copy of 'susanne.tar.Z' to your home system, get the material into a usable state by the successive commands:

```
uncompress susanne.tar.Z
```

and:

```
tar -xf susanne.tar
```

If you are not a Unix user, you need to retrieve the other version of the Corpus, which will be uncompressed using the PKUNZIP software on an IBM-PC. First, set ftp transfer mode to binary by typing the command:

```
bin
```

at the ftp prompt. Then retrieve the appropriate version of the Corpus by typing:

```
get susanne.zip
```

Having transferred a copy of the Corpus to your home machine, uncompress it with the command:

```
pkunzip -x susanne.zip
```

In either case (whether you have followed the Unix or the non-Unix instructions) you should now have the Corpus split up into its 65 files, one of which, 'SUSANNE.doc', is a text file describing the format and contents of the 64 data files.

To log out of the ftp connexion, type:

```
bye
```

If you encounter any problems, please send an e-mail message to archive@black.ox.ac.uk or archive@uk.ac.oxford.vax.

The Bergen Corpus of London Teenager Language (COLT)

Anna-Brita Stenström and Leiv Egil Breivik
University of Bergen

Thanks to a generous grant from the Norwegian Research Council we are now in the process of collecting a corpus specially designed for teenager talk. We are aiming at a corpus of half a million words of the English spoken by London teenagers aged 13 to 17.

In order to get the largest possible social spread we have selected five different London districts, each of which may be considered representative of a particular social grouping.

The technique adopted for the recording is modelled on that employed for the collection of the British National Corpus. In our case, boys and girls from each London district act as recruits; they carry a small Walkman which makes it possible to record any conversation they are engaged in with (preferably) youngsters of the same age for a period of two to five days. They have also been instructed to insert all relevant details, such as who speaks when, where, and with whom, etc, in a conversation log.

Since we regard COLT as a complement to the British National Corpus, where teenager language is not specifically aimed at, we have decided to use an identical transcription scheme, ie a broad orthographic transcription with very little prosodic information and no phonetic marking, but where the transcriber concentrates on speaker turns, including overlapping and interrupted speech, pauses, laughter, voice quality, and so on. The recordings will be divided up into 'conversations', and speaker turns will be analysed in terms of 'sentence-like' units, identified by punctuation marks.

The initial stage, which involves gathering and transcribing data, started in March this year and will run over one year, and we hope that some of the material will be available for research by the end of the year.

Thirteenth ICAME Conferance, 3-7 June 1992

Christian Mair
University of Freiburg

The 13th ICAME conference – competently organised by Jan Aarts, Pieter de Haan, Nelleke Oostdijk and helpers – took place at the Hotel 'de Plasmolen' from 3 to 7 June 1992. A combination of largely asparagus-based fare, receptions and excursions – to historic Nijmegen and the Kröller-Müller collection at Otterloo – provided participants with the energy and drive necessary to keep up with a stimulating and demanding conference programme. To accommodate a large number of contributors, events were structured into four categories – 'long' (40-minute) and 'short' (20-minute) presentations, posters (with presenters allowed a five-minute introductory talk), and 'hands-on' demonstrations of software. This made for a tight schedule, but participants generally felt that the alternative – streaming presentations into parallel thematic sections – would not have been a good idea, one of the advantages of ICAME after all being that it is a meeting ground for 'mechanics' and 'drivers', i.e. people who know how a computer-readable corpus and associated software work and people who merely use corpora to get from point A to B in some philological line of enquiry. As someone tending toward the 'driver' end of the spectrum I freely admit that I still do not precisely know what 'simulated annealing' is but sitting through a number of papers on parsing techniques certainly helped me to get closer to an answer. Conversely, the kind of nice distinctions in the data unearthed by word-loving philologists in their scrutiny of data will show the more computer-minded worker how far he will have to go to meet the needs of some consumers.

If in the following I fail to mention some contributions, this is certainly no verdict on their quality. I simply want to give a survey, necessarily subjective, of what I perceived as the major areas of emphasis in current work on English computer corpora.

As usual, a number of presentations reported on new projects or progress in ongoing ones. Sidney Greenbaum, having presided over an ICE (International Corpus of English) workshop immediately preceding the ICAME Conference, reported on this project. Of the matching one-million-word corpora documenting British, American, Canadian, Aus-

tralian, New Zealand, Caribbean and several Asian and African Englishes, the English component will be available shortly in tagged form. Of the proposed fringe projects within ICE, only Sylviane Granger's corpus of learner English is currently making progress. Gavin Burnage and Dominic Dunlop reported on the British National Corpus, 90,000 000 words of written and 10,000 000 words of spoken English, planned to be usable by 1994. Alex Fang (Hong Kong) gave a talk on tagging his institution's 1,000 000-word corpus of computer science language. 'Electronic English' is considered and documented as a new variety by M. Collot and Nancy Belmore. Among the smaller projects presented in various stages of completeness many had a diachronic orientation, making one wonder about the justification of the letter M in ICAME.

Thus, Susan Wright and Josef Schmied focussed on the importance of sociolinguistic and contextual determinants in their two Early Modern English corpora. Merja Kytö introduced the audience to a half-million word appendix to the Helsinki Corpus containing colonial American texts. Christian Mair's proposed 1991 replica of LOB (press sections A, B, C completed to date) addresses more recent changes in British English. Other projects presented, such as Gerry Knowles' database of spoken English or Ian Lancashire's Renaissance dictionaries corpus, eventually to become part of a Renaissance knowledge base, are probably best described not as traditional corpora but as ventures into hypertext.

Also in a category of its own is the 'monitor' corpus compiled at the Research and Development Unit for English Language Studies in Birmingham. Susan Blackwell talked about how to efficiently clean up the masses of data fed through the monitor while Antoinette Renouf presented filters designed not only to dredge up new word forms but, a more difficult but linguistically far more rewarding project, also new combinations of words, existing words used in new meanings and contexts and shifts in frequency of use.

The technical problems of parsing were at the centre of a considerable number of contributions. Clive Souter started off with a survey of resources. Kees Koster (Nijmegen), Akiva Quinn (ICE, London), Eric Atwell and Robert Pocock (Leeds), and Ted Briscoe (Cambridge) talked about their respective projects. Other presentations with a computer-science orientation ranged from Elizabeth Eyes' and Geoff Leech's surveys of recent work within UCREL at Lancaster, comprising all aspects of corpus annotation, through Louise Guthrie's and Jim Cowie's work on automatic lexical disambiguation to a diverse array of corpus utilities concentrated in the demonstration room. All were interesting, but it

seems fair to single out Knut Hofland's and his co-workers' ICAME CD-ROM as this year's landmark achievement.

An exhaustive treatment of the linguistically orientated contributions is similarly impossible. Again, the selective mention of four examples is not meant as an indictment against the ones passed over, but as an indication of the breadth of topics covered. Graeme Kennedy (Wellington) made us think about what precisely it is that we mean when we use *when* when we talk. Christine Johansson's poster was one of several contributions dealing with aspects of relative clauses, which – for some reason I am not completely aware of – seem to be the corpus linguist's most favoured grammatical construction. How corpora could be used to complement and/or demolish the case for prescriptive grammar could be gleaned from Pam Peters' poster on whether to split or to not split the infinitive. And although meant as mere illustration, Antoinette Renouf's list of 'new words of February 1991' as based on the *Times* will find their way into many a classroom.

I personally benefitted much from a 'fringe' event, namely the discussion of the role of corpora in teaching which was inspired by Steve Fligelstone. I hope for similar such forums at least at irregular intervals at future ICAME gatherings.

Dominic Dunlop and Gavin Burnage may have wondered about a few sniggers when they put one of their transparencies on the projector. It contained the word 'acadedmic,' with a misspelling which if it was not intended was almost Freudian. ICAME 13, however, was not dead but very much alive.

In 1993, ICAME will meet in Zürich, Udo Fries and Gunnel Tottie being the hosts.

Note

A volume of papers from the Thirteenth ICAME Conference is now available: Jan Aarts, Pieter de Haan, and Nelleke Oostdijk (eds), *English language corpora: Design, analysis, and exploitation.* Amsterdam & Atlanta, Ga.: Ropodi.

The First International Colloquium on English Diachronic Corpora (St Catharine's College Cambridge, 25-27 March, 1993)

Merja Kytö and Matti Rissanen
University of Helsinki

Susan Wright
University of Cambridge

In recent years, the interest in the compilation of corpora containing texts from the earlier periods of English has increased rapidly, together with the development of new methods and aids for tagging and parsing the texts in these corpora. The number of computer-assisted studies of the history of English has soared and there are important major research projects making effective use of databases of early English.

Last year, at the ICAME Conference in Nijmegen, some participants interested principally in diachronic corpora came to the conclusion that it was time to improve the contacts and cooperation between scholars who are active in collecting corpora or, in other ways, combining computer-assisted methods with the study of the history of English. It was felt that the expertise and know-how in many departments and research centres all over the world could be put to more effective use. It was also felt that consideration ought to be given to the problems of duplication of work and waste of energy, and the dangers of the disintegration of the rapidly developing field of corpus-based research of the history of English.

As a result of these considerations, the authors of this report invited a small group of specialists to the First International Colloquium of English Diachronic Corpora, which was held at St Catharine's College, Cambridge, in late March, 1993. Twenty-six scholars from eight countries (Austria, Canada, Finland, Germany, Great Britain, Norway, Switzerland and the United States) attended, along with nine Cambridge observers.

The two-day colloquium concentrated on introductions of the work in progress in English historical corpora, thesaurus, atlas and dictionary projects, and software development. Time was also reserved for discus-

sions of various problems in corpus compilation and management. Among the topics discussed, the questions of lemmatization and normalization, and genre or text type coding aroused a lively exchange of opinions. The applicability of the TEI (Text Encoding Initiative) to the coding of early English texts was discussed on the basis of an introduction by Professor Stig Johansson (Oslo University).

The following corpus projects were introduced.[1]

1. Innsbruck Computer Archive of Middle English Texts (Manfred Markus, University of Innsbruck)

The ICAMET project will produce a large corpus of complete Middle English prose texts. The present size of the corpus is close to two million words.

2. The Cambridge-Leeds Corpus of Early Modern English (Jonathan Hope, University of Leeds, and Susan Wright, University of Cambridge).

This project will produce a corpus of Early Modern English texts, dating from c. 1600 to c. 1800. The first phase of the project focuses on the texts written by London-based collaborative playwrights of the early 17th century and on early 18th century writings, currently 900,000 words in electronic form. (For an introduction, see the forthcoming Papers from ICAME 13.)

3. The Century of Prose Corpus (Louis T. Milic, Cleveland State University)

This corpus of c. 500,000 words of literary and non-literary British English, dating from 1680 to 1780, has been available since 1990. (For an introduction, see *ICAME Journal* 14, 1990, 26-39.) The corpus is being revised and the documentation volume will be published soon.

4. The Zurich Corpus of English Newspapers (Udo Fries, University of Zurich)

The ZEN corpus will contain texts taken from British newspapers from 1660 to the inception of *The Times* newspaper.

5. The diachronic part of the Helsinki Corpus of English Texts (Matti Rissanen, University of Helsinki)

The Helsinki Corpus, which consists of c. 1,5 million words of texts dating from c. 750 to c. 1710, has been available since 1991. (For an introduction, see *ICAME Journal* 16, 1992, 7-27, and the *Manual*

compiled by M. Kytö, 1991). A volume with a detailed introduction to the corpus and pilot studies will come out in 1993.

6. The Helsinki Corpus of Older Scots (Anneli Meurman-Solin, University of Helsinki)

This 600,000-word corpus is a supplement to the Helsinki Corpus introduced above. It contains samples from 64 Scots texts dating from 1450 to 1700. The corpus will probably be available in a year's time.

7. The Corpus of Early American English (Merja Kytö, University of Helsinki)

This supplementary corpus to the Helsinki Corpus will consist of half-a-million words of texts, written in New England and Southern colonies in the 17th and early 18th century.

8. The Lampeter Corpus of Early Modern English Pamphlets (Josef Schmied, Bayreuth University)

This corpus will consist of a sample of 80 pamphlets of variable length (between 2,000 and 12,000 words), dating from 1640 to 1740. All the texts are found in the Tract Collection at the Old Library of Saint David's University College, Lampeter, Wales. The project will finish in 1994. At the moment, a third of the texts exist in electronic form.

9. A Corpus of Nineteenth-Century Letters (David Denison, University of Manchester)

This corpus, primarily intended to support the writing of the Syntax chapter of Vol. IV of the *Cambridge History of the English Language*, consists of letters dated between 1861 and 1918 (currently c. 100,000 words). It may be enlarged, time and funding permitting, if its utility is proved.

10. A Representative Corpus of Historical English Registers (Douglas Biber, Northern Arizona University, and Edward Finegan, University of Southern California)

The ARCHER Corpus consists of British and American English texts and covers the period from 1650 to 1990, divided into 50-year periods. The samples are taken from seven written and four spoken genres or registers. The target sampling is ten texts, at least 2,000 words per genre, in each 50-year period, of both British and American English. So far, most of the samples have been collected, edited and auto-tagged; well over a half have also been tag-edited.

The current state of the following computer-assisted thesaurus, atlas and dictionary projects were reported.

1. The Historical Thesaurus of English (Christian Kay, University of Glasgow) and the Old English Thesaurus (Jane Roberts, King's College London)

The Historical Thesaurus will be completed in a few years' time. The archive of data, comprising around 700,000 meanings is virtually complete, and about 80% of the material has been organized under the heads of a classification based on the material, mental and social worlds. Over 70 major categories, comprising some 225,000 records, have been entered in the database.

The Old English Thesaurus will serve as a pilot study for the Historical Thesaurus. The inclusion of the Old English material in the Historical Thesaurus will, for the first time, present evidence for the Anglo-Saxon vocabulary obsolete by 1150 alongside with the forms that replaced them. Once the pilot thesaurus is completed, the compilation of an annotated Thesaurus of Old English will start at King's College London.

2. Linguistic Atlas for Early Middle English and Linguistic Atlas for Older Scots (Margaret Laing, The Institute of Historical Dialectology, University of Edinburgh)

The LAEME and LAOS projects will produce linguistic atlases based on exhaustive corpora of medieval texts. The words are tagged both for the meaning (PresE translation or Old English form) and for word class and syntactic function. The date, location and 'type' of the texts are also coded. There is a program to produce chronological charts of the forms of an item.

So far, the projects have been operating with 'prototype' tagging programs. These programs are currently being reviewed and some are being revised or rewritten, and it is hoped that an improved tagging program will be produced in the near future. The scholar in charge of the LAOS database is Dr Keith Williamson of the Institute of Historical Dialectology.

3. An Early Modern English Dictionary Database (Ian Lancashire, University of Toronto)

The purpose of this project is to compile a dictionary based on 35 bilingual and monolingual English dictionaries dating from 1500 to 1660. The textbase to be collected will hold about ten million words. The dictionary will give equivalents in English, Latin, French, Italian, Spanish

and other languages, comment on their usage and contextualize them in illustrative phrases and sentences. Using appropriate software, English words 'hidden' in the explanations of non-English terms in bilingual dictionaries can be made word-entries and the foreign-language lemmas can be used in 'explaining' words.

4. The Johnson Dictionary (Anne McDermott, University of Birmingham)

This project will produce a new edition of Johnson's Dictionary. It will mainly be based on the first and fourth editions of the Dictionary. Particular attention will be paid to the sources of the text examples illustrating the meanings of the words.

Dr Jeremy Smith illustrated the use of concordances, dictionaries and thesauruses with reference to the dialectal strata of the vocabulary of the late-fourteenth-century poet John Gower.

Two software presentations were included in the program of the colloquium. Knut Hofland (University of Bergen) introduced the CD-ROM 'ICAME Collection of English Language Corpora', which contains the Brown, LOB, London-Lund, Kolhapur and Helsinki Corpora, and demonstrated the use of the WordCruncher and TACT concordance programs with these corpora. He also gave a presentation on various distribution lists and file servers currently available in international networks. Professor Raymond Hickey (Bayreuth University) introduced the LEXA corpus processing software system, created by him for personal computer use. The set of programs in LEXA will carry out lexical analysis and information retrieval tasks. It has been particularly developed to be used with diachronic material, but the general nature of the software permits its application to any set of texts. The software, with the 3-volume manual, can be ordered from the Norwegian Computing Centre for the Humanities in Bergen.

A follow-up workshop will be arranged in connection with the ICAME Conference in Zürich in May 1993. The proceedings of the Colloquium, with more detailed information on the items mentioned above, will come out later this year.

The English Department of the University of Helsinki will be responsible for collecting information in the fields covered by the Colloquium. An information sheet will be distributed to interested scholars and departments regularly in printed form and via electronic mail. Information on all developments in historical English corpora, software adapted to be used with these corpora, or major research projects making active use of computer-assisted methods, will be gratefully received by Prof. Matti

Rissanen, Department of English, P.O.Box 4 (Hallitusk. 11) FIN-00014 University of Helsinki, Finland.

Note

1. The name of the Colloquium participant introducing the project (in most cases the compiler of the corpus or the project leader) is given in brackets.

ICAME services

The CORPORA distribution list

Knut Hofland
Norwegian Computing Centre for the Humanities

The CORPORA list is open for information and questions on text corpora such as availability, aspects of the compilation and use of corpora, software, tagging, parsing, bibliography, etc. The list currently (April 1993) has about 500 members.

To join the list send a message to LISTSERV@UIB.NO with the line 'SUB CORPORA' firstname lastname in the body of the letter.

NB! This is not a full LISTSERV, but only a reduced one to handle subscriptions to distribution lists. If you want to get log files etc, send a message to FILESERV@HD.UIB.NO with subject HELP.

To contribute to the list, send messages to CORPORA@HD.UIB.NO

Messages to this address will automatically be resent to all the members on the list.

PLEASE note the difference between the addresses:

LISTSERV@UIB.NO	subscription
CORPORA@HD.UIB.NO	messages to everybody on the list
FILESERV@HD.UIB.NO	file server

Other correspondence should be sent to the list administrator:

Knut Hofland
Norwegian Computing Centre for the Humanities,
Harald Haarfagres gt. 31,
N-5007 Bergen, Norway

Phone: +47 5 212954/5/6, Fax: +47 5 322656,
E-mail: knut@x400.hd.uib.no

ICAME file servers

Knut Hofland
Norwegian Computing Centre for the Humanities

FILESERV

The machine nora.hd.uib.no has been established as a mail-based server for the Norwegian Computing Centre for the Humanities. Information is grouped in different directories, some of which have information in Norwegian only.

Some of the available directories:

corpora	Information from the distribution list CORPORA, log files, etc.
icame	International Computer Archive of Modern English
info	Information on texts, projects etc., mostly in English
konferanser	Information on conferences, mostly in English
mac	Macintosh programs
ncch	Norwegian Computing Centre for the Humanities Information in English
nettinfo	Information on network resources, mostly in English
pc	MS-DOS programs
unix	Unix programs

The server is called FILESERV and runs the DECWRL archive server. FILESERV accepts three types of commands; several commands can be placed in the body of the mail message. However, the results will be sent in one file, so do not request several large files in one message. The commands are:

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

Example: If you want to get the index for the CORPORA and the KONFERANSER directories and the file log.started.920918 in the CORPORA directory, send the following two notes

('index' and 'send' commands cannot be put in the same message, the 'send' commands will then be ignored):

To: fileserv@hd.uib.no
Subject: whatever

index corpora
index konferanser

To: fileserv@hd.uib.no
Subject: whatever

send corpora log.started.920918

FTP SERVER

The files are also available via anonymous FTP from nora.hd.uib.no (129.177.24.42). To make use of this server, you must have access to a machine connected to Internet with TCP/IP and a program running the FTP protocol.

Example: To get the directories of the server write the following:

ftp nora.hd.uib.no
anonymous
your e-mail address
cd pub
dir

The server has a directory for uploading; this is writeable but not readable.

cd incoming
(binary) (if transfer of programs or 8-bit data)
put xx-program.zip

Please send a note and a description to knut@x400.hd.uib.no if you upload any files!

Other commands:

get <file>	
mget <dir-mask>	(to get several files, example: mget *.ex)
cd <directory>	(change directory)
cd ..	(up one level in the directory tree)
binary	(set binary transfer, for transfer of programs or 8-bit files)
ascii	(set transfer of 7-bit text data)

GOPHER SERVER

The information is now also available through our Gopher server at nora.hd.uib.no (port 70). If you are connected to the Internet (with TCP/IP protocol), you can get client versions of Gopher for MS-DOS, Macintosh and Unix. Gopher is a tree-structured menu system, and several hundred servers are connected.

Main menu on the nora.hd.uib.no machine:

Internet Gopher Information Client v1.02

Root gopher server: nora.hd.uib.no

→ 1. About this Gopher at NCCH
2. Andre Gopher tjenere (other Gopher servers)
3. Corpora (distribution list)
4. Forskjellig (various) Info
5. Humanistisk datasenter (NCCH)
6. ICAME (Text corpora)
7. Konferanser (Conferences)
8. NCCH file servers
9. Nettverk (Network) Info
10. Nordic Linguistic Bulletin
11. Norwegian Computing Centre for Humanities
12. Programs

Press ? for Help, q to Quit, u to go up a menu.

Questions about these services can be directed to:

Knut Hofland,
Norwegian Computing Centre for the Humanities,
Harald Haarfagres gt. 31,
N-5007 Bergen, Norway

Phone: +47 5 212954/5/6 Fax: +47 5 322656
E-mail: knut@x400.hd.uib.no

Texts available through ICAME

The following corpora are currently available through the International Computer Archive of Modern English (ICAME). For information on the CD-ROM, see further p. 145.

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

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

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

Brown Corpus, other versions (diskette and CD-ROM): See p. 145. The WordCruncher version is described in an article by Randall Jones, *ICAME Journal* 11, pp. 44-47.

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

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

LOB Corpus, tagged version, horizontal format (tape, diskette, and CD-ROM): 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 (tape and CD-ROM): Each word is on a separate line, together with its tag, a reference

number, and some additional information (indicating whether the word is part of a heading, a naming expression, a quotation, etc).

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

LOB Corpus, other versions (diskette and CD-ROM): See p. 145.

Lancaster Parsed Corpus (tape and diskette): This corpus consists of syntactically analysed sentences from each text category of the LOB Corpus, amounting altogether to over 133,000 words. See the presentation by Geoffrey Leech in *ICAME Journal* 16, pp. 124-126.

London-Lund Corpus, complete text (computer tape, diskette, and CD-ROM): The London-Lund Corpus contains samples of educated spoken British English, in orthographic transcription with detailed prosodic marking. It consists of 100 'texts', each of some 5,000 running words. The text categories represented are spontaneous conversation, spontaneous commentary, spontaneous and prepared oration, etc. The original version of the London-Lund Corpus (87 texts) is no longer available. As regards the versions available, see p. 145.

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

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

London-Lund Corpus, supplement (tape and diskette): The 13 texts not included in the original version of the London-Lund Corpus. See the presentation by Sidney Greenbaum, *ICAME Journal* 14, pp. 108-110.

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

Kolhapur Corpus, original version (tape, diskette, and CD-ROM): A

million-word corpus of printed Indian English texts. See the article by S.V. Shastri, *ICAME Journal* 12, pp. 15-26.

Kolhapur Corpus, other versions (diskette and CD-ROM): See p. 145.

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

Polytechnic of Wales Corpus (tape or diskette): Orthographic transcriptions of some 61,000 words of child language data. The corpus is parsed according to Hallidayan systemic-functional grammar. There is no prosodic information. See further *ICAME Journal* 13 (1989), p. 20ff, and 15 (1991), pp. 55-62.

Helsinki Corpus (tape, diskette, and CD-ROM): A selection of texts covering the Old, Middle, and Early Modern English periods, totalling 1.5 million words. See the article by Merja Kytö and Matti Rissanen in *ICAME Journal* 16, pp. 7-27. As regards the versions available, see p. 145.

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

There are available printed manuals for the LOB Corpus (the original manual and a supplementary manual for the tagged version), the Helsinki Corpus, and the London-Lund Corpus. Printed manuals for the Brown Corpus cannot be obtained from Bergen. Users of the London-Lund material are also recommended to consult J. Svartvik (ed.). *The London-Lund Corpus: Description and Research*, Lund University Press, 1990.

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

Programs available through ICAME

Together with the diskettes or tapes with texts we include some freeware programs. With the PC versions we include TACT, a text indexing and retrieval program developed at the University of Toronto. With the Mac versions we include a Hypercard stack and Free Text Browser, for indexing and text retrieval.

With Unix tapes we include an indexer/browser in C code and also the HUM package, for producing word lists and concordances. These programs are also available from our file servers. We collect freeware programs from different sites and make them available through our file servers (or information on how to get the programs from other sites).

We also distribute the Lexa program (see the article by Raymond Hickey in this issue) and the index/view version of WordCruncher; see the order forms accompanying this journal. As regards programs distributed with the CD-ROM, see below.

The ICAME CD-ROM

The ICAME Collection of English Language Corpora is a new CD-ROM produced and distributed by the Norwegian Computing Centre for the Humanities. It includes the following corpora (for some information on these corpora, see pp. 142–143):

Brown Corpus: Bergen text version I and II, for MS-DOS, Macintosh and Unix. A modified Bergen version II indexed by WordCruncher 4.4 and TACT for MS-DOS and Free Text Browser for Macintosh.

LOB Corpus: Tagged and untagged original text versions, for MS-DOS, Macintosh and Unix. A tagged horizontal version indexed by WordCruncher 4.4 and TACT for MS-DOS and Free Text Browser for Macintosh.

Kolhapur Corpus: Text version for MS-DOS, Macintosh and Unix. A version indexed by WordCruncher 4.4 for MS-DOS.

London-Lund Corpus: Original text version for MS-DOS, Macintosh and Unix. An edited version indexed by WordCruncher 4.4 and TACT for MS-DOS and Free Text Browser for Macintosh.

Helsinki Corpus: Text version for MS-DOS, Macintosh and Unix. 1-file,

3-file and 11-file versions indexed by WordCruncher 4.4 and TACT for MS-DOS.

As the material is provided in a number of versions, it should be easy to use. The following programs are distributed with the disc: WordCruncher View, TACT, and Free Text Browser.

The disc contains a number of information files, including full lists of texts for the Brown, LOB, and Kolhapur corpora, and the list of speakers for the London-Lund Corpus. It also contains information on network resources, such as discussion lists and sites for downloading of programs, Netnews, lists of electronic text projects and some linguistic freeware programs. Manuals for the Helsinki and London-Lund corpora are distributed with the disc. See further the order form accompanying this journal.

Conditions on the use of ICAME corpus material

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. Publications making use of the material should include a reference to the relevant corpus (or corpora), giving the name of the corpus and the distributor.

Information for contributors

Language. All contributions should be in English. Contributors whose native language is not English should have their manuscripts gone through by a native speaker before submission.

Format. Contributions should preferably be submitted as ASCII files on diskette, together with a printout made from your word-processing system. As regards other possible formats, consult the editors before submission of your manuscript.

Headings. The title of the paper should be followed by the author's name and academic affiliation. Sections and sub-sections should be numbered. Headings should **not** be singled out typographically (by boldface, capitalization, or the like).

Tables and figures should be numbered and titled. They should always be referred to by their number, **not** by expressions like 'see the diagram below' or 'in the following table'. Tables should be submitted in a separate file. Drawings, graphs, and other illustrations must be reproducible originals.

Quotations. Use single quotation marks, except for quotes within quotes. Long quotations should be indented and given without quotation marks.

Examples should normally be numbered and set apart from the text following standard linguistic practice. Short examples in the running text (words or phrases) should be underlined.

Notes should be placed at the end of the paper. References to notes in the text should be indicated as follows: *1, *2, etc.

References should conform to standard linguistic practice. References in the text should follow this pattern: Francis (1979: 110) defines a corpus as 'a collection of texts assumed to be representative of a given language, dialect, or other subset of a language, to be used for linguistic analysis'. The list of references at the end of the paper should be presented as shown by these examples:

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

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

Renouf, Antoinette. 1987. Corpus development. In *Looking up: An*

account of the COBUILD Project in lexical computing, ed. by J. M. Sinclair. 1-40. London & Glasgow: Collins ELT.

Tottie, Gunnel, and Ingegerd Bäcklund (eds.). 1986. *English in speech and writing: A symposium*. Studia Anglistica Upsaliensia 60. Stockholm: Almqvist & Wiksell.

Authors should be given with their full first names, unless they always use the initials themselves.

Reviews. The heading of a review should contain the information shown in the following example:

Roger Garside, Geoffrey Leech, and Geoffrey Sampson (eds.). *The computational analysis of English: A corpus-based approach*. London: Longman, 1987. 196 pp. ISBN 0-582-29149-6. Reviewed by Gunnel Källgren, University of Stockholm.

Review articles should have a title, followed by the author's name and affiliation, and the information on the book(s) reviewed, as shown above.

Submission, books for review. Contributions, as well as books for review, should be sent to one of the editors:

Stig Johansson
Department of British
and American Studies
University of Oslo
P.O. Box 1003
Blindern
N-0315 Oslo 3
Norway
E-mail: stigj@ulrik.uio.no

Anna-Brita Stenström
Department of English
University of Bergen
Sydnesplass 9
N-5007 Bergen
Norway
stenstroem@hf.uib.no

The editors are grateful for any information or documentation which is relevant to the field of concern of ICAME.

ICAME Journal is published by the Norwegian Computing Centre
for the Humanities (Humanistisk datasenter)
Address: Harald Hårfagres gate 31, N-5007 Bergen, Norway.
Telephone: +47 5 212954 Telefax: +47 5 322656
E-mail: icame@hd.uib.no
ISSN 0801-5775