12. Concordancing and Practical Grammar

Tony Jappy

This paper illustrates one way in which the computer can be used to complement and exploit a theoretical course in English grammar. The practical application of grammatical knowledge in the computer-assisted analysis of various genre categories offers a macroscopic, or "bird's eye," view of the texts in the corpus. The work allows us to test assumptions concerning the linguistic structure of given types of discourse. After a brief review of relevant aspects of the English verb phrase, the paper discusses the methodological problems of concordancing and offers a simple methodology for analyzing results. Finally, the paper shows how valid, and in certain cases, surprising conclusions can be drawn from this form of macroscopic discourse analysis.

Introduction

The purpose of the present study is to describe how computer-assisted concordancing can be taken from the field of research and be put to pedagogical use in a TESOL environment. This is now possible with the general availability in the classroom of sophisticated computational technology, together with relatively user-friendly commercial text retrieval programs such as MicroOCP, MicroConcord, TACT and Wordcruncher.[1] The paper will examine the distribution of the English verb phrase in a commercially available corpus and take the reader step by step through all the stages of one form of pedagogically-oriented concordancing.

To this end, the study first shows how the verbal forms of English relate to two ways of constructing propositions; it then illustrates how concordance programs can retrieve the appropriate linguistic data from an interesting and manageable corpus; finally, it discusses the problem of processing and harmonizing the data returned in the searches and gives a simple statistical method of obtaining and interpreting information from these and similar data. It will be seen that, although obtained in a practical teaching environment, the results yield interesting insights into the nature of the distribution of verb forms over the genre categories investigated, thereby contributing to the students' increased grammatical awareness.

The Subject-Predicate Relation in Discourse

Since Aristotle, the proposition has been considered the basic item of information in discourse.[2] One of the most fundamental linguistic operations involved in the production of propositions is the association, by the speaker, of a subject (S) with a predicate (P). This, it will be shown, determines the equally fundamental distinction made within the English verb phrase between the so-called 'contiguous,' or simple forms, as in *I see, I saw*, etc., where the relation between subject and predicate is direct and immediate, and the 'non-contiguous,' compound forms, where the relation between subject and predicate is mediated by various types of auxiliary and combinations thereof, e.g. *I have seen, I was seen, I didn't see, I might have been seen*, etc. This distinction then constitutes the basis of our study of the macroscopic distribution of the two distinct 'patterns' of predication, S_P and S_AUX_P, and the various verbal forms which realize them, across selected genres in this particular corpus of contemporary English.

In order to comprehend fully the formal distinctions utilized in the concordances, and to understand the objective or subjective values that can be attributed to them, consider the following sets of example utterances, which all conform to the S_P pattern:

(A1) I came, I saw, I conquered.
(A2) Lear walks to the front of the stage, bows to the Fool,...
(A3) Lineker runs down the left wing... dribbles past a defender... gets his cross in...
(A4) Gary Lineker plays for Tottenham.
(A5) My friend Jack eats sugar lumps.
(A6) Your train leaves at nine-fifteen.

Utterances (A1-3) are examples of narration, i.e. of the representation of events in sequence, whether in narrative, stage directions or sports commentary.[3] In (A4-6), on the other hand, the simple verb forms are used, not to narrate events in sequence, but to characterize the subject of the predication in various ways. Discourse characterized by these forms, which can conveniently be subsumed under the general term 'reporting,' tends to be factual, objective and 'positive' in the sense that it represents only what is the case.

Fortunately, linguistic representation in English is not restricted to such

forms, and utterances (B1-8) below variously exhibit the expression of the speaker's subjectivity. These examples fall within the broad linguistic categories of aspect, the passive voice, mood and modality, and illustrate the S_AUX_P pattern:

(B1) Mrs. Thatcher is visiting Zambia.
(B2) Gale-force winds have caused havoc all across the continent.
(B3) High winds and heavy seas have been causing further problems in the southern part of Britain.
(B4) Passengers were led to safety after a fire broke out on the London Underground.
(B5) My study doesn't have a bar.
(B6) PET DOG MAY TRAP KILLER
(B7) Rain will spread from the west.
(B8) John should have been digging the garden.

Utterances (B1-3) illustrate the realizations of aspect in English, by which we mean the various ways in which the speaker represents the degree of completion of a process with respect to some reference point. Since selecting a reference point and using it to evaluate the degree of completion of a process are discursive strategies, and not features of the referential world, it follows that the various aspectual markers of English (***have + en, be + ing***, and their combined form) to be found in an utterance are traces of the speaker's involvement in the utterance and not a feature of the situation being represented. In (B4), by contrast, speaker involvement takes the form of a radical change of sentence perspective, in which the object of the process functions both as subject and theme of the clause. Since there are obviously no passive events in the referential world, it follows that any change in Subject Verb Object (SVO) perspective in English can only be a discursive strategy, and, in the resultant passive voice, the subject and predicate are mediated by the marker ***be + en***. Similarly, with indications of mood: to put matters crudely, as there simply is no such thing as a negative or interrogative situation, it follows that any negative or interrogative elements in an utterance can only have been introduced by the speaker. Typically, but not exclusively, these negative and interrogative elements are carried, as in (B5), by ***do*** used as an auxiliary.[4] Finally, the expression of irrealis mood, and with it, degrees of the speaker's evaluation of the validity of the subject-

predicate relation, is a form of subjective appreciation; in this way (B6) a headline from a tabloid newspaper, (B7) a 'prediction' from a weather forecast issued by the BBC, and (B8) a counterfactual statement (he should have been, but he wasn't), all illustrate yet another form of speaker-involvement in the utterance.[5] In short, utterances (A1-7), realizing the S_P pattern, give the impression of an objective, positive report of the events represented or of individual participants therein. Utterances (B1-8), on the other hand, realize the S_AUX_P pattern and, in various ways, express the speaker's subjective, often explanatory, evaluation of the situation or state of affairs being referred to (cf. Hopper, 1979: 217).

We note, finally, that the present and past tenses are common to both sets of utterances, and function as signs of assertion, i.e. of the speaker's acceptance of responsibility for the proposition s/he is advancing. It thus follows that AUX, in the S_AUX_P pattern, can be expanded to **(modal)** ***(have + en) (be + ing) (be + en) / (do)***, where the parentheses indicate that the item is optional; and that the distribution of ***do*** is parallel to that of the other auxiliaries. Since this complex predicative pattern is positively marked morphosyntactically by the auxiliary forms discussed above, the programming of a suitable concordancer to retrieve its various realizations from a set of texts provides TESOL students with an interesting exercise in applied grammar.

Method

The first task is to establish frequency counts for the linguistic features we happen to be interested in, here the forms of the English verb (minus the modals). We conduct the searches using Micro-OCP and the Lancaster University-IBM UK Spoken English Corpus. This corpus of relatively formal spoken English dating mainly from the mid-80s is principally based upon recorded material from the BBC, runs to some 52,000 tokens,[6] and comes in various guises (all in ASCII format). Note that in a TESOL environment, the orthographic version should be used in preference to one of the tagged[7] versions available. Clearly, the latter would render the grammatical analyses more trustworthy. However, experience has shown that the unnatural appearance of parsed corpora could lead to pedagogical disaster, and the texts are best edited with a wordprocessor to insert the appropriate referencing conventions.

These reference conventions show our computer program (in this case Micro-OCP) where the different genres begin and end, and also allow us to

identify to the program the presence in the texts of simple present and preterit forms: concordance programs have no way of telling whether the token *works*, for example, is a plural noun or the third person of the verb, or whether the token *worked* is a preterit or a past participle. While such editing is a painstaking task, it should not be forgotten that a project of this sort is a useful class exercise and that the instructor has at his disposal an unlimited supply of manual labor... Given below is an extract from one of the twelve Radio Commentary texts showing the referencing conventions, where <C Comment> indicates that the current genre category is Commentary and where the symbols % and & have been arbitrarily chosen to identify the verbal forms as the simple present and preterit respectively:[8]

> *<C Comment>*
> The New York Times correspondent &looked out of his car window, and &told me the guerrillas had taken Suchitoto: did I want a ride? I &jumped in, and we &set off at the manic speed which, for some reason, %is a characteristic of the way all journalists &drive here in El Salvador. Suchitoto %is a particularly bullet-holed, bombed-out town; a tenuous government stronghold in the heart of guerrilla controlled territory, thirty miles north of the capital, San Salvador. Every correspondent here %agrees that the final six mile stretch through Suchitoto %is the most eerie, the scariest bit of road in El Salvador. The last reporter to be killed here, back in March, was shot dead in a crossfire on that road; another reporter's car &hit a mine there two years ago, and he too was killed; everyone's had narrow scrapes on the Suchitoto road. There &were three of us in the car, all rather nervous. The third reporter &was with the Washington Post: a war correspondent for twenty years who'd covered Vietnam. The Washington Post man &said he &hoped that, at an army checkpoint just before the final stretch to Suchitoto, they would stop us from going through. They didn't. [...]

There are various types of proprietary concordance programs available, but in the present study we restrict our attention to Micro-OCP. Although a pre-Windows software package, it can be clicked up from the Windows File Manager, is menu-driven by means of the function keys, and, on present-day

486 machines, will process the sort of program illustrated below across the small corpus described here in a matter of seconds. Moreover, unlike TACT and Wordcruncher, the program works directly on text files.[9] Given below is one of the suite of programs that are needed to retrieve all the verbal forms under scrutiny. This particular program 'trawls' through the corpus for the regular and irregular plural forms of the English present perfect (e.g. not only *have worked*, but also *have heard, have written, have caught, have come*, etc., plus such contracted negative forms as *haven't come*, etc.):[10]

Program 1

```
*input references cocoa "<"to">".
        comments between "["to"]".
*action do concordance
        include only phrases "have*d", "have* *n", "have* *e",
        "have* *k", "have* *ng".
        references C = 5. Contexts sorted by references.
*format layout length 78 and lines 0 below entries.
        references right. headwords left same line. titles
        "have*.ins" left on line 1.
*go
```

Concordancers generally work by matching patterns supplied by the user with the words found in the corpus. Our patterns are indicated in the *include only phrases* line in the program above, where the * symbol is a wild-card matching any (possibly null) sequence of characters following the stem *have*. Obviously, it is possible to program Micro-OCP to produce a concordance based simply upon the string *have*, but the disadvantages are, firstly, that the student is not required to think about the variety of patterns associated with the irregular present perfect forms and, secondly, the results would need a considerable amount of editing to weed out inappropriate items. As it is, the second entry in the concordance[11] below has to be discounted as the *have* *t* pattern has 'pulled in' a verb + noun group and not a present perfect:

Table 1. Extract from the concordance produced by Program 1

have*.ins		
eir selection. They mus	**have brought** this lot by the foot. I can't	Ficti
rain - northern areas will	*have right intervals*	*News*
tted dark. His eyes must	**have burst**, he thought, they were full of	Ficti
these economic ideas	**have cast** their spell	Lectu
before, it wouldn't	**have caused** the stir t	Lectu
rain and gale-force winds	**have caused** havoc all	News
Jews. And now attitudes	**have changed:** Germans are going to	Comm
why God should	**have chosen** to create	Lectu
nobody could	**have conceived** the man	Magaz
says Mr. Powell's remarks	**have dashed** government hopes that the	News
entation - the smaller parties	**have done** unexpectedly well	News
said the lion. I could easily	**have eaten** him, only I'd promised you.	Ficti
one hundred thousand	**have gone**, and West Germany	Comm
learn Turkish, housewives	**have got** together to help draw up the	Comm
families who lived in them	**have left**, taking advantage of a double	Comm
ough fine Gael and Labor	**have lost** ground to the opposition, it's	News

Concordance-based searches of this sort,[12] which have to be processed by the students in an exercise euphemistically referred to as 'hand editing,' i.e. manual counting, yield the sort of results given on Table 2:[13]

(continued overleaf)

Table 2. Frequency counts for eleven linguistic features:

Category:	*Comm.*	*News*	*Lect.*	*Mag.*	*Fict.*	*Total*
Tokens:	9066	5235	11922	4710	7299	38232
simple pres	164	57	139	61	44	465
preterit	175	81	182	88	522	1048
present perf	71	74	44	29	25	243
past perf	25	18	29	5	29	106
present cont	31	6	8	6	21	92
past cont	12	5	5	3	44	69
passive	28	38	15	5	1	87
do*(n't)	9	0	12	2	23	46
did(n't)	12	3	6	14	17	52

The next stage is to make sense of the widely differing genre lengths and the differences in the totals of the observed frequencies returned per verbal feature. As matters stand in Table 2, it is impossible to compare the frequency counts for the preterit and passive features, for example, as they have different totals (1,048 and 87 tokens respectively), or, say, the frequency counts for all features for the fiction and magazine genres, as these do not have the same number of tokens. Finally, since this is a course for students majoring in English as a foreign language, the statistical treatment of the data needs to be relatively simple. These problems can be resolved by adopting a simple strategy advanced by the French statistician, Michel Volle (1974).

Volle's preoccupation was with the often voluminous and unwieldy tables with which statisticians are obliged to work when writing reports. He suggested that the tables themselves be relegated either to the official publications from which they were extracted or to an appendix at the end of the statistician's report and that only the most 'informative' cells in the table be discussed. He proposed a very simple way of identifying such cells, which amounts to adapting the well-known test for significance, the chi-squared test. Once significance is computed, Volle identifies the informative cells as those 'partial' chi-square cell values which contribute the greatest percentage of 'information' to the

chi-squared total. Thus, in a 20 by 20 contingency table, for example, eleven cells out of the possible four hundred might contribute, say, 65% of the information in the table, in which case the statistician would restrict his attention to these. This method is adapted to identify the greatest value per feature from the raw frequencies given on Table 2, and consists of the following stages:

First, one calculates what the expected figure would be if the average distribution of each structure in each genre were equal. This is achieved by multiplying the total number of occurrences of a given form by the total number of tokens for each given genre category and dividing the result by the total number of tokens in the corpus:

total tokens of **form**	x	total tokens in **genre category**	÷	total tokens in **corpus**

For example, there are **243** tokens for the present perfect in the corpus. In the genre *Commentary* there are **9066** tokens. The total number of tokens in the corpus is **38232:** 243 x 9,066 = 2,203,038 ÷ 38,232 = 57.6. In other words if all genres had equal representation of each grammatical structure, one would expect 57.6 occurrences of present perfect in the *Commentary*. In the example below, the features are the present and past perfect forms, and the symbols O and E represent, respectively, the observed and expected frequencies:

Table 3. Computing the expected frequency (E) per cell

Cat.	*Commentary*		*News*		*Lecture*		*Magazine*		*Fiction*	
Tokens	9066		5235		11922		4710		7299	
	O	E	O	E	O	E	O	E	O	E
present perf	71	57.6	74	33.3	44	75.8	29	29.9	25	46.4
past perf	25	25.1	18	14.5	29	33.1	5	13.1	29	20.2

Second, one uses the chi-squared value to calculate how far the actual occurrence of the observed structure differs from the expected distribution. Each cell is therefore computed from the formula $(O-E)^2/E$. For example as

calculated above, the expected number for the present perfect in the *Commentary* was 57.6 compared to the observed frequency of 71. Following the formula, we obtain: 71-57.6 = 13.4 x 13.4 = 179.56 ÷ 57.6 = 3.1. As before, the example uses the present and past perfect features:

Table 4. Computing chi-square

Cat.	*Comm.*	*News*	*Lect.*	*Mag.*	*Fict.*	*chi-sq.*
present perf	3.1	49.7	13.3	0.02	9.9	76
past perf	0	0.8	0.5	5	3.8	10.1

Third, in order to extract the maximum amount of information from the calculation of the test, each cell's contribution to the chi-squared total is expressed as a percentage of that total (Table 5). For example, if we add together all the figures in Table 4 for the present perfect, we get a total of 76 (3.1 + 49.7, etc). The figure of 49.7 for the genre *News* in present perfect is 65% of the total (49.7 ÷ 76 x 100 = 65%). The following example is limited to the present perfect, since the chi-squared scores for the past perfect turn out to be insignificant and therefore do not appear in subsequent tables:

Table 5. The results obtained from computing 'Volle' scores per present perfect cell

Cat.	*Comm.*	*News*	*Lect.*	*Mag.*	*Fict.*
pres perf	+	+65%	-18%	-	-13%

Table 5 shows that for the present perfect feature, the *News* category scores heavily (65% of the 'information' contributing to the chi-squared total) while the *Lecture* and *Fiction* categories exhibit significant deficits as far as this particular feature is concerned.

Fourth, the method is applied to all the the data returned by the concordances as they appear in Table 2, and the scores compared.[14] Note that, for the sake of simplicity, the results given in Table 6 arbitrarily include only cells contributing at least 10% of the 'information.'[15] One must calculate signifi-

cance per row and not for the total number of cells in the table; in other words, Table 6 is a compilation of ten different sub-tables:

Table 6. 'Volle' scores for cells contributing at least 10% of the information to the total.

Category:	*Comm.*	*News*	*Lecture*	*Mag.*	*Fiction*	*chi-sq.*
simple pres	+52%	-	-	+	-45%	49.8
preterit	-	+	-10%	-	+80%	644.0
present perf	+	+65%	-18%	-	-13%	76.0
present cont	+11%	+40%	-41%	-	+	36.3
past cont	-	-	-14%	-	+78%	91.5
passive	+	+69%	-	-	-18%	83.0
do*(n't)	-	-20%	-	-	+72%	32.0
did(n't)	-	-11%	-28%	+39%	+22%	22.9

Thus, frequency counts converted into percentages of this sort are readily comparable. However, since the chi-squared scores pertain only to the data set out in Table 2, the disadvantage is that any conclusions drawn are inevitably corpus-specific.

Finally, we round the percentages on a scale from 1 to 10, with minus values omitted as displayed on Table 7:

(continued overleaf)

Table 7. Volle scores simplified to show the single most important cell per feature:

Category	*Co*	*Ne*	*Lec*	*Ma*	*Fi*
simple	mm	ws	t.-	g	ct
present	5	-	-	+	-
preterit	-	+	-	-	8
pres perfect	+	7	-	-	-
pres contin	+	4	-	-	+
past contin	-	-	-	-	8
passive	+	7	-	-	-
do*(n't)	-	-	-	-	7
did(n't)	-	-		4	+

Discussion

Table 7 is obviously a very simple 'rule of thumb' representation of the relation between linguistic feature and genre in this particular corpus, but it nevertheless yields highly suggestive results encouraging the student to reflect upon the compatibility of the values of the verbal forms with the genres with which they are associated in the corpus. If we consider the fiction genre, for example, we find that fiction in this corpus tends to favor past forms.[16] Furthermore, there is a very marked compatibility between this genre, the subjective value of past imperfective aspect (the past form of **be** + **ing**, here +8), and the event-oriented, objective nature of the preterit (+8). This appears to confirm Hopper's statement (1979:216) that imperfective aspect has a backgrounding, commentative function with respect to the foregrounding function of narration. There is, however, a noticeable incompatibility between the fiction genre and the use of the passive in its narrative function (-18% on Table 6), suggesting that the change of sentence perspective that the narrator operates by means of the passive is less 'natural' in narrative than the thematically more consistent use of SVO order.[17]

Table 7 also shows that three of the five genres are positively characterized by the features under investigation and, although obtained in a teaching project and not a full-scale piece of research, this fact raises two interesting

theoretical issues. Firstly, as mentioned above, the *Fiction* genre is characterized by the objectivity of the simple past and the subjectivity of imperfective aspect. Secondly, and no less interestingly, the *Radio News* genre obtains a high rating for the subjective nature of the present perfect (+7), with the resultative value it derives from the speaker's relating the consequence of some past event to the moment of broadcast, and an even higher rating for the clause-level speaker manipulation exhibited by the passive (+7), an index of the news editor's preoccupation with the victims of such events and concern for thematic continuity within each news item being presented. No doubt the BBC would be chagrined to learn that its news broadcasts are less than objective, but the fact of the matter is that the examples to be found in the SEC corpus display two features that indicate considerable manipulation of the linguistic medium by the speaker, and to that extent are subjective in the sense ascribed above to the S_AUX_P pattern. Interestingly, Biber characterizes his dimension 5 as seeming to "mark informational discourse that is abstract, technical and formal versus other types of discourse" (1988:112-113). In other words, discourse obtaining high scores for such features as passives with or without an explicit agent will generally be abstract, technical and informal. Typically, this is the dimension of supposedly objective scientific reports. However, as the data provided by the news broadcasts show, any departure from SVO word order with its attendant change of sentence perspective is a subjective rather than objective discourse manoeuvre. For reasons of cohesion, the report-writer is in fact imposing his own perspective-seeking subjectivity upon the reports given of the events in question.

Conclusions

Firstly, as a class project this 'bird's-eye' view of texts and the verbal forms that compose it can yield insights into the functions of the English verb. Moreover, the results tend to confirm initial assumptions concerning the objective or subjective nature of the categories involved. As expected, the most flexible of the genres, fiction, shows the greatest range of values.

Secondly, while the figures presented in Tables 2, 6 and, 7 obviously have only a relative, corpus-specific value, unlike the 'absolute' value of Biber's descriptive statistics, they tend to confirm the meanings generally attributed to the features investigated, and illustrate a simple method for the investigation of other linguistic properties of corpora.

Thirdly, certain compatibilities and incompatibilities to be found in corpora would no doubt be more thoroughly processed with a specialized text-processing statistical package. However, as such packages generally begin by lemmatizing[18] texts before processing them, much of the information discussed above would be lost, and, as any non-statistician will appreciate, such a program is more appropriate to a research laboratory than to a TESOL environment, where the techniques described above represent the upper limit in statistical complexity.

Tony Jappy studied modern languages at Oxford. He is an associate professor in English linguistics at the University of Perpignan. His research interests and publications deal with iconicity theory and metaphor, the relation between semiotics and linguistics, and computing and linguistics. He has been teaching courses in computing since 1985.

Notes:

1. Micro-OCP and its less versatile but more user-friendly stablemate, MicroConcord, can be obtained from the Oxford University Press. Tact is bundled with the Lancaster University-IBM UK corpus described below and obtainable from the Norwegian Computer Centre for the Humanities, Harald Harfagres gate 31, N-5007 Bergen, Norway. For Wordcruncher, currently shipping in a Windows version, contact Johnston & Company, PO Box 6627, Bloomington, IN 47407, USA. For the reader wishing to take the macroscopic analysis of texts further, I would recommend Butler, 1992 as a good introduction to the logistics of the problem, and Biber, 1988, to which I make frequent reference in the text, as an outstanding example of the study of linguistic variation across corpora.

2. It is self-evident that, unlike the proposition, *John is a student* neither the expression *is a student* nor *John is a* yields information.

3. Care must be taken with the term "commentary" for the simple verb forms of sports commentary are, in fact, examples of narration, and must not be confused with the metalinguistic term "commentary." The idea of distributing verbal forms within a narration-commentary "dimension" was first mooted by Weinrich 1973. The distinction

was taken up and developed by Paul Hopper in a series of articles, principally Hopper, 1979, and now seems generally established.

4. Note that mood can be realized by any of the auxiliaries, and is not the exclusive preserve of *do*.

5. The discourse in which such forms are found tends to be factual and objective: as the French linguist Henri Adamczewski put it, in such cases the presence of the speaker is not coded (1982: 41). The non-contiguous forms, on the other hand, all exhibit various facets of the speaker's involvement in the utterance. Obviously, limitations of space preclude a more exhaustive exposition of the theoretical issues involved in the interpretation of the various forms of the verb phrase. For this, the reader is referred to the discussion and references to be found in the article by Rémi Lapaire and Wilfrid Rotgé in Volume 2.

6. A token can be described as the actual word in the text. If one had the following in a span of text: *comes, coming, comes, come, come*, one would have five different tokens for the type *come*. On the other hand, one might have two tokens for two different types:

   He *heads* for home as fast as he can.
   Tonight, *heads* will roll.

7. A tagged version might label each verb, for example, according to tense and aspect.

8. In the present case, the analyses have been restricted to the five major genre categories in the corpus: Radio Commentary, Radio News, Radio Lectures, Radio Magazines and Radio Fiction, totaling approximately 38,000 tokens.

9. Prior to any analysis, TACT for example, requires the construction of a text database. MicroConcord, on the other hand, will work not only with ASCII files but also with those produced with most of the major word-processors.

10. Obviously, the search pattern could be made more general by including only phrases such as "ha* *d" instead of writing separate programs for "has* *d," "have* *d" or "had* *d" etc.: the programming would be simplified, but the editing of the output file would be fastidious to say the least, since it would include phrases such as *"(his) hands moved," "hardly touched,"* for example, which both conform to such a search pattern. Note that each instruction must be terminated by a period.

11. Called a KWIC concordance, it displays keywords in context. In the example, the keywords have context to the left and right, are highlighted in bold, and the genres they belong to are indicated by their first five letters (cf. line 6 in the program) on the right-hand side.

12. To obtain a list of all occurrences of the simple present and the preterit, one would normally program Micro-OCP to produce, not a concordance, but a word list of just those words beginning with % and & respectively. In this case, the Micro-OCP **words* option (not needed in the program given above) would have to specify that these characters are 'additional' letters of the English alphabet.

13. As with the tagging of the simple present and preterit forms, such a task is daunting to the individual, but as this is a class project, the work can be divided up conveniently and parceled out to the advantage of one and all.

14. The same remark applies here. Obviously, the results should be cross-checked, but with a class of volunteers, this should not present a problem.

15. Significance was calculated at <0.001 for 4df.

16. Note that the present perfect obtains a negative value (-13 on Table 6), which is surprising as perfective aspect is one of Biber's principal indices of narrative concern.

17. This surmise is confirmed by the figures for third person subject pronouns (+8) and hapax legomena (-4), which were not included in the chart. The pronoun value suggests a higher degree of topic-subject continuity within narrative episodes than in the other genres, and the high negative hapax value suggests correspondingly less lexical variation, a rough index of narrative and thematic unity.

18. To lemmatize a word is to find its dictionary form.