Automated Text Analysis: Cautionary Tales

C. N. BALL
Georgetown University, Washington, USA

Abstract

The increasing availability of electronic text and text analysis tools has made it possible to analyse vast amounts of data in a short amount of time. However, natural language processing is not a solved problem, and even large research systems representing decades of development do not perform at the level of human language processors. Since such systems are not sufficiently robust for general use, most literary and linguistic corpus analysts make use of heuristics and simple tools for text analysis. But while such 'shallow' approaches offer improvements in speed and accuracy over traditional manual methods, there are many pitfalls for the unwary. In this paper we consider some pitfalls and temptations that attend the automated analysis of large text corpora: sample size, the recall problem, analysing only what is easy to find, and counting what is easiest to count. We suggest that, given the state of the art in text processing tools, such tools must be used with a full awareness of their limitations, and should be coupled with or replaced by manual methods when appropriate.

1. Introduction

The increasing availability of on-line text corpora and text analysis tools has made it possible to analyse vast amounts of data in a short amount of time. In Biber's (1988) important study of the Lancaster-Oslo-Bergen (LOB) and London-Lund (LL) corpora, for example, the frequency of sixty-seven linguistic features is analysed in 481 texts, and Biber observes: 'To analyze this number of texts without the aid of computational tools would take several years' (Biber, 1988, p. 65). However, automated text analysis (or more generally, natural language processing) is not a solved problem, and recent tests of large research systems on unrestricted text revealed varying degrees of accuracy (Sundheim, 1991). Since most literary and linguistic corpus-analysts are not using such broad-coverage systems, but are rather relying on heuristics and small tools, accuracy is clearly an issue. Furthermore, known limitations of home-grown algorithms may lead to the exclusion of some phenomena from study, while the relative ease of obtaining some statistics (e.g. word count) may result in the use of problematic frequency metrics. In this paper, we review some of the pitfalls of automated text analysis, with particular reference to large corpora, and suggest that given the present state of the art, automated methods and manual methods for text analysis must go hand in hand.

2. Corpus Analysis and Small Samples

Corpus design is clearly a crucial issue for the corpus linguist or literary analyst: the results of the analysis hold true for the corpus, and can only be generalized to the extent that the corpus is a representative sample. When working with an on-line corpus that has been designed for general research needs, it is vital to consider the sampling methods employed and whether the samples are likely to be representative with respect to the phenomena under investigation. As a simple example, consider the problem of tracking the frequency of *wh*-forms in restrictive relative clauses in spoken and written Early Modern English (1500–1700). The most suitable electronic corpus for investigation of this problem is the Helsinki Diachronic Corpus (Nevalainen and Raumolin-Brunberg, 1989), but the spoken data consists of sampled transcripts of five trial proceedings (1554, 1600, 1603, and 1685). Given the fact that restrictive *wh*-relatives are relatively rare during this period (Rydén, 1966), it is clearly preferable to analyse a larger sample by hand than to draw conclusions from a small electronic sample which is unlikely to be representative.

3. The Recall Problem: You Don't Know What You're Missing

Once the issue of corpus design has been addressed, the electronic text analyst will proceed to the selection and use of tools for text analysis, including concordances, taggers, parsers, and search utilities.[1] Here there are many pitfalls for the unwary, the most serious of which can be characterized as 'the recall problem'. Recall and precision are measures of retrieval effectiveness generally used in information retrieval studies, where precision is the proportion of retrieved material that is relevant, and recall is the proportion of relevant information that was retrieved (Salton, 1989, p. 284). Suppose that the analyst uses an automated search procedure to retrieve tokens of a specified type, e.g. relative clauses. It is not difficult to judge the precision of the search: it will be obvious to the trained analyst when a token is retrieved that contains an interrogative clause rather than a relative clause, for example. Poor precision can be dealt with by hand, e.g. by editing the output of the search to eliminate non-instances of the type. However, precision errors often lead to a narrowing of the search criteria, which in turn may lead to a decrease in recall.

The danger lies in the difficulty of assessing recall in a large corpus: it is generally impossible for the analyst to know what has been missed without analysing the entire corpus by hand. In an important study of retrieval effectiveness in a large commercial full-text retrieval system, Blair and Maron (1985) found that users were much more confident in the accuracy of the system than was warranted: users were satisfied with the system's

Correspondence: Catherine N. Ball, Department of Linguistics, Georgetown University, Washington, DC 20057, USA. E-mail: cball@guvax.georgetown.edu

performance, when in fact less than half of relevant information was retrieved. Users are able to judge the precision of their searches, but are generally unable to judge what has been missed in a large database. There is an important lesson here for corpus analysts, particularly where what is being investigated is a complex linguistic phenomenon.

To illustrate this point, I have culled two examples of simple heuristics from the literature, which were used to retrieve linguistic phenomena as the basis for frequency analysis. The authors are to be commended for publishing their algorithms: it is more common in reports of corpus-based research for the search method to be left unspecified. The first example is a set of patterns from Biber (1988) for finding zero complementizers in subordinate clauses in the LOB and LL corpora.

(1) Search criteria for subordinator-*that* deletion (Biber, 1988, p. 244)

(1a) PUB/PRV/SUA + (T#) + demonstrative pro/ SUBJPRO

(1b) PUB/PRV/SUA + PRO/N + AUX/V

(1c) PUB/PRV/SUA + ADJ/ADV/DET/POSSPRO + (ADJ) + N + AUX/V

These patterns appear to state that there is a zero complementizer when a verb of a certain class is immediately followed by a subject noun phrase and an auxiliary or a main verb. However, the algorithm does not allow for the possibility of other clause-initial constituents such as prepositional phrases and parentheticals, and it encodes an overly simplistic view of NP structure (ignoring *wh*-clauses, co-ordination, post-nominal modifiers, recursion, etc). In fact, NP structure cannot be represented by a finite set of patterns of this type. The following examples from the LOB Corpus[2] illustrate a few of the tokens that would not have been retrieved by the search criteria as stated:

(2a) LOB A19 153: I think what I'm really seeking all the time is the source of Original Sin in myself.

(2b) LOB E19 112 . . . I think the Santa Rita or Cabernet would match this food.

(2c) LOB A29 122: I think the way it is going on is very worrying, but nothing more.

An algorithm with greater recall but much less precision is given by Rissanen (1991):

> The problem of tracing zeroes must, in this case, be solved in an indirect way, by checking all occurrences of all verbs which take an object clause with *that*, in order to find all the possible instances of zero. In this way, collecting a 'complete' set of instances of *that* and zero from the entire Helsinki Corpus and printing it into workable text files is a matter of a few days' hard work. (Rissanen, 1991, p. 275)

As Rissanen observes, this approach may not yield perfect recall, as there is the possibility that some verbs in the corpus occur only with zero. Still, it appears to be relatively safe. The problem for a large study like Biber's, which included many linguistic features, is that a few days to complete each search may amount to year's worth of sustained effort. On the other hand, a frequency study based on an inaccurate search of the data cannot be expected to yield reliable results. Great caution should be exercised in the formulation of heuristics for retrieving tokens of linguistic phenomena from a corpus. Heuristics should be reported along with the findings, and should be treated with skepticism by the reader.

4. Confined by the Limitations of Text Analysis Tools

Traditional text analysis can be characterized as a search method that has perfect recall but no precision at all: everything in the corpus is presented for analysis by the human brain. It is a rich and rewarding experience, in which the analyst brings to bear a wealth of knowledge about language and the world, and acquires new knowledge and new research directions in the process. To read a corpus is to encounter the unexpected and the undocumented at every turn. There is much, in fact, to be said in favour of poor precision in linguistic analysis: with perfect precision, we find exactly what we said we were looking for, and no more.

Unfortunately, mental text analysis is attended by tedium, errors, and the passage of time. Not all text is equally engaging; not everything that is read is attended to, and what *is* noticed must be coded and transferred to electronic form for further analysis. In the perfect world, text analysis would be totally automated, while the human analyst reads for insight and new hypotheses. Unfortunately, there is currently no software that captures the linguistic and non-linguistic knowledge of the experienced linguist (or even the adult native speaker). The state of the art for processing unrestricted text is a collection of text processing tools which can augment, but not replace, the human analysis process. Excessive reliance on these imperfect computational aids may constrain the analysis process in undesirable ways. In this section, I shall consider some implications of measuring only that which is easy to count, and excluding that which is difficult to find automatically.

4.1 *Finding Nothing*

There are many possibilities in English for the zero realization of constituents, including zero complementizes [(2a–c) above]. Some others are illustrated below.

(3) Zero relatives
LOB K06 45: . . . we should have someone ∅ we could trust.

(4) Gapping
LOB E24 117 For some children will be needed a stimulus, for a domineering few ∅ the gentle brake.

(5) Verb-phrase deletion
LL 1 2a 1611370 1 1 B: . . . I'm very sorry I cannot teach at the institute . . . I will do my best to find someone who can ∅ . . .

(6) Absolute possessives
G33 150: . . . the phrase recurs throughout the writings of Count Henri Patrick Marie Russell-

Killough. The latter's ∅ was an affectionate and generous character.

(7) Telegraphic style: zero subjects, objects, prepositions, determiners . . .

(7a) LOB G12 149: My diary for the 18th records: ∅ Turned out ∅ 7:30 and dressed in plain clothes. During breakfast ∅ got a signal from ∅ C.-in-C. . . .

(7b) LOB E01 143: First launder ∅ crochet and then pin ∅ to the required shape, ensuring that all lines of the crochet are accurate.

In a sense, ∅ is a member of many paradigms. But no concordance utility or string search can find instances of nothing in an untagged corpus, and as Rissanen (1991) observes, even tagged corpora do not generally tag empty constituents. The identification of non-overt elements requires manual effort, a parsed corpus, or a robust parser, but there are considerable accuracy and coverage issues with currently available parsers, and the number of parsed corpora is small.[3] There are at least two problems for electronic corpus analysis that thus arise.

The first is that the analyst may simply not consider ∅ as a possibility. As a simple illustration, consider the problem of identifying and counting clauses in a corpus. The mental parser will register a gapping such as that in (4) as a clause.[4] But when designing a heuristic to identify clauses automatically, will the designer remember gapping? Francis and Kucera (1980, p. 550ff.), in their careful discussion of an algorithm for identifying clauses, seem to have overlooked this possibility. Since gapping is a highly marked syntactic construction (Levin, 1987), however, its omission probably does not affect their results for the Brown Corpus.

A second possibility is that the analyst may deliberately exclude such phenomena from consideration, which may seriously affect the validity of any frequency analysis of the larger class. A case in point is Biber (1988), where zero relatives [as in (3)] are omitted from the analysis of relative clauses because they cannot be identified reliably by automatic means (Biber, 1988, p. 221). But zero relatives are, unlike gapping, a central grammatical phenomenon. Their exclusion means that the total number of relative clauses cannot be determined, with consequences that I shall discuss in Section 5 below.

4.2. *Frequency Metrics*

In general, the limitations of text analysis tools may lead the analyst to focus on that which is easy to identify and count automatically. In terms of frequency metrics, for example, relative frequency of syntactic phenomena in printed texts is typically expressed in terms of number of tokens per clause-level unit, e.g. restrictive relative clauses as a percentage of all relative clauses (Beaman, 1984), clefts per 100 clauses (Ball, 1991), or preposings per 100 independent clauses (Ward, 1988). However, in studies which use large electronic corpora, relative frequency is often measured in terms of number of tokens per *N* words (Biber, 1986; 1988; 1992; Biber and Finegan, 1988; Collins, 1991, *inter alia*). The reason, presumably, is that it is a relatively simple matter to calculate the number of words in a large electronic corpus,[5] but to count clauses requires a reliable heuristic, a parsed corpus, or considerable manual effort. Whether word count is a valid frequency metric is an issue that has not been addressed in the literature, but if it should turn out to be unreliable, the results of a number of word-based frequency studies will clearly be called into question. In this section I shall explore this issue, using Collins' (1991) study of *it*-clefts in the London-Lund and LOB corpora. The conclusions extend to all word-based frequency studies.

The *it*-cleft construction, exemplified in (8a–b) below, is difficult to identify automatically with good precision. Collins used a method with poor precision but very good recall, the essential ingredient being a search for all tokens with *it* (Collins, 1991, p. 25). Clefts were then culled from the results.

(8a) LOB G07 38: It may have been then that Trelawny contrived to do his copying. (= Collins 1991 (25))

(8b) LOB F03 175: It's their interest you want—not their sympathy. (= Collins 1991 (64))

Collins counted the number of words in each text group in the two corpora, and used word count to produce a frequency metric of clefts per 10,000 words. Text groups were then ranked by relative frequency, with results such as those shown in Table 1 below.

From such results, Collins draws inferences about factors that affect the relative frequency of clefts: for example, the ordering in Table 1 is seen to represent a 'factual/descriptive' versus 'opinionative/rhetorical' dimension, with clefts being more frequent in the more persuasive text groups (Collins, 1991, pp. 186–7). The reliability of these results depends on the accuracy of the search, the accuracy of the word count, and the reliability of the frequency metric. I shall consider only the latter.

In a word-based frequency analysis, to say that a phenomenon occurs with equal relative frequency in two samples is to say that equal amounts of text, measured in words, will yield the same number of tokens. But relative frequency should be a measure of the number of times something occurs within the number of opportunities for it to occur. Put differently, if we are measuring X within Y, each Y should be an opportunity for an X to occur (where X is a member of the class Y). Thus, we can measure the number of three-letter words within all words; the number of present tense verbs within all tensed verbs, the number of interrogative clauses within all clauses, and so on. Choosing the appropriate class is, of course, crucial and requires considerable thought. But it requires little reflection to see that the appropriate unit of measurement for relative frequency of cleft sentences and other clause-level syntactic phenomena is not the word: it is not the case that every word provides an opportunity for a cleft sentence to occur. Clefts and words are not members of the same class.

This fundamental observation is forced upon us when constructing a chi-square matrix. For example, the dif-

Table 1 Frequency of clefts in the informative categories of LOB (from Collins, 1991, Table 7.5)

Text category	Clefts	Words	Frequency per 10K words
H (Miscellaneous documents, reports, etc.)	11	60600	1.8
A (Press: reportage)	24	88543	2.7
J (Learned and scientific writings)	82	161389	5.1
E (Skills, trades, and hobbies)	39	76567	5.1
F (Popular lore)	55	88685	6.3
G (Belles lettres, biography, essays)	104	155109	6.8
D (Religion)	23	34226	6.8
B (Press: editorial)	39	54294	7.0
C (Press: reviews)	26	34216	7.6

Table 2 Distribution of clefts in the informative categories of LOB

Text category	Clefts	Non-clefts	Total
H (Miscellaneous documents, reports, etc.)	11	?	60600
J (Learned and scientific writings)	82	?	161389
	93	?	221989

rerence between the relative frequency of clefts in text groups H and J appears quite large (1.8 versus 5.1), but is the distribution of clefts significant? To answer this sort of question the chi-square test is commonly employed, requiring a matrix of the type shown in Table 2. Here the problem is apparent: we have a total measured in words and a count of cleft sentences. The second column should be the complement of the first column (i.e. non-clefts), but then the total must be number of clefts plus the number of non-clefts, which will be expressed as number of clauses, independent clauses, or however we have decided to count non-clefts. In short, the chi-square matrix as it stands is ill-formed, word count is apparently useless, and we require a different sort of count.

If word count is not appropriate for measuring the relative frequency of syntactic phenomena in a corpus, what unit *is* appropriate? Opportunities for a speaker/writer to choose a clausal construction occur minimally at the level of the clause, but we might also consider the independent clause (cf. Ward, 1988, p. 94). The question now arises: if we are given only word count, is it possible to convert from word count to number of clauses or number of independent clauses, and thereby provide a meaningful interpretation for word-based frequency studies? In general, this does not appear to be the case: we cannot predict the number of clauses in a text given only number of words.

What is required is some constant X such that (9) is true:

(9) word count/X = number of clauses
X = word count/number of clauses

X will thus be the mean length of clauses measured in words, or the word/clause ratio.[6] The crucial issue is not so much the value of X, but whether it is a constant. It never been claimed, to my knowledge, that there is such a constant, and there is in fact considerable evidence that the word/clause ratio is variable across texts.

Let us begin with mean sentence length, one of the earliest phenomena investigated in connection with differences between spoken and written data. Gibson *et al.*, 1966, in a study of forty-five American college students, compared 750–1000 word essays and five-minute extemporaneous speeches on the same topic. The data were analysed on five measures, including average sentence length. The results for the latter, shown below, were found to be highly significant ($P < 0.01$).

It is thus shown that the ratio of words to sentences is variable across samples of spoken and written data. But the sentence is too large a unit for our purposes: in extreme cases, a text may consist of only a single sentence (Crystal and Davy, 1969, p. 197), while containing many clauses. More useful is the notion of independent clause. In a follow-on study of differences between spoken and written data, O'Donnell (1974) analysed data in terms of mean length of independent clauses. These are referred to as *T-units*, but the definition makes it clear that the notion is equivalent to the independent clause: '[A] T-unit . . . contains one independent clause and the dependent clauses (if any) syntactically related to it. In traditional grammatical terms, it can be

Table 3 Mean sentence length in spoken and written data (from Gibson, 1966, Table 1)

Variable	Speeches	Essays	*P*
Mean sentence length	16.28	18.18	<0.01

Table 4 Mean T-unit length in spoken and written data (from O'Donnell, 1974, Table 1)

Variable	Interview	Newspaper
Mean T-unit length	17.92	24.97

the equivalent of a simple sentence or a complex sentence . . .' (O'Donnell, 1974, p. 103).

The data for O'Donnell's study were produced by one educated adult American male. Two samples of 100 independent clauses each were drawn from the subject's newspaper columns and a radio interview, with the results shown in Table 4.

These results again show a difference between spoken and written data, and are thus consistent with the findings of Gibson *et al.* (1966). But there is additional information we can extract from the two studies: because one sentence may contain more but not less than one independent clause, or T-unit, we can infer that mean T-unit length must be less than or equal to mean sentence length. Combining the results of the two studies, we can see that the ratio of words to independent sentences is variable not only between spoken and written data, but across all the texts (Table 5).

If the independent clause is chosen as the appropriate unit of measurement for relative frequency of clefts, we may stop here, having shown that we cannot compute number of independent clauses given only word count. However, clefts occur in both independent and dependent clauses, so the clause appears to be a more suitable unit of measurement. A relevant study of mean clause length in the Brown Corpus is reported in Francis and Kucera (1980), with the results shown in Table 6 below.

As can be seen, there is a fairly wide range of variation in mean clause length across text groups in the Brown Corpus, from Romance (5.49) to Miscellaneous (8.59). Given the differences in sentence length and T-unit length reported between speech and writing, it may be expected that we could find spoken genres with even lower scores than Romance. These results are for American English, but the observation of variability should carry over to the London-Lund and LOB corpora of British English.[7] It is clear that the ratio of words to clauses is not constant across text groups, and that word-based frequency analyses of clause-level phenomena are therefore not equivalent to clause-based frequency analyses.

Summary

A number of studies of linguistic phenomena have been conducted using word count as the basis for measuring relative frequency. However, the denominator in a frequency ratio must be of the same class as the numerator, and syntactic constructions such as the cleft sentence are not words, but clauses. Their relative frequency should therefore be measured in terms of number of clauses. If the ratio of words to higher-level units such as the clause were constant across texts, word-based frequencies could be converted to equivalent clause-based frequencies. However, studies show that in general, this ratio is not constant. Number of clauses is not predictable from word count alone, nor is the number of independent clauses or sentences. Because the ratio of words to higher-level units is variable, no conclusions about the distribution of syntactic phenomena within a corpus can be drawn from word-based frequency studies, and such studies are non-comparable. While words in an electronic corpus are relatively simple to count, word count does not yield interpretable results for syntactic phenomena.

5. Danger: Hidden Variables

In the section above, we showed that relative frequency of linguistic phenomena must be measured with respect to the appropriate class. In the case of constructions, the relevant class may be the clause. The number and

Table 5 Mean T-unit length, spoken and written data

Variable	Speeches	Interview	Essays	Newspaper
Mean T-unit length	≤16.28	17.92	≤18.18	24.97

Table 6 Mean sentence and clause length in the Brown Corpus (Francis and Kucera, 1980, Table 6.1)

	Genre	Words per Sentence	Predications per Sentence	Words per Predication
P	Romance and Love Story	13.41	2.44	5.49
L	Mystery and Detective	12.62	2.29	5.50
N	Adventure and Western	12.75	2.30	5.54
K	General Fiction	13.82	2.41	5.74
M	Science Fiction	12.94	2.23	5.81
R	Humor	17.52	2.82	6.21
E	Skills and Hobbies	18.53	2.59	7.15
B	Press: Editorials	19.66	2.73	7.21
F	Popular Lore	20.26	2.81	7.22
G	Belles Lettres	21.35	2.93	7.29
D	Religion	21.21	2.90	7.33
J	Learned	22.31	2.84	7.85
A	Press: Reportage	20.27	2.63	7.88
C	Press: Reviews	21.06	2.62	8.03
H	Miscellaneous	24.07	2.80	8.59

length of clauses is thus a factor not controlled for in word-based frequency analyses: a hidden variable. A similar problem arises in the case of frequency analyses of relative clauses, but here there are many more known variables to be controlled for. To illustrate the effects of hidden variables, we shall consider Biber's (1988) quantitative study of the LL and LOB corpora, which included the following types of restrictive relative clause (with Biber's examples):

(10) Relatives (Biber, 1988, p. 234ff., features 29–34)

(10a) *that*-relatives on subject position
The dog that bit me.

(10b) *wh*-relatives on subject position
The man who likes popcorn.

(10c) *that*-relatives on object position
The dog that I saw.

(10d) *wh*-relatives on object position
The man who Sally likes.

(10e) pied-piping
... the manner in which he was told.

Mean word-based frequencies for each of these types were calculated for each of the text groups within the corpora, and the scores were included in a factor analysis. Now, this approach assumes that the choice of relative marker (e.g. *that* vs. a *wh*-pronoun) is conditioned only by grammatical role and text type. But clearly, not every word in a text is a potential relative marker: it is within the class of relative clauses that the opportunity to choose a particular relative marker occurs, and any two text groups may exhibit differing relative frequencies of relative clauses. Indeed, Biber cites Beaman's (1984) study, in which it is shown that spoken and written Pear narratives (Chafe, 1980) differ in frequency of relative clauses. This fact alone confounds Biber's analysis, and renders the results problematic.

In light of the many studies of relative markers which measure frequency within the class of relative clauses, it may be asked why Biber chose word count as a frequency metric. Part of the answer may lie in the difficulty of identifying all relative clauses automatically. Consider the subclass of direct object restrictive relatives, with relative frequencies drawn from Quirk's (1957) study of relative markers in educated spoken British English:

Table 7 Relative markers for direct object restrictive relatives (data from Quirk, 1968, p. 104)

wh	*that*	zero	Total
83 (22%)	145 (39%)	148 (39%)	376

A major member of the class is the Ø-relative, absent from Biber's analysis because of the difficulty of identifying tokens automatically (Biber, 1988, p. 221, cf. Section 4.1 above). But without counting Ø-relatives, it is not possible to calculate a total for the class of object relatives, and hence impossible to measure relative frequency of any member within the class.

There are other variables which must be controlled for in a quantitative study involving relative markers,[8] and here I will simply mention one that is even more difficult to identify automatically: type of antecedent. It has been widely recognized, at least since Quirk (1957), that the choice of restrictive relative marker in standard English is affected not only by grammatical role (which Biber has taken into account), but also by type of antecedent. Thus, for example, in a restrictive relative clause with a human-like ('personal') antecedent, where the subject of the clause has been relativized, *who* is vastly preferred to *that*, while with non-personal antecedents, *that* and Ø are generally preferred to *which*.

(11) Personal subject relative
LOB K02 21: This rather surprised me from a young man who was otherwise so sophisticated.

(12) Non-personal relatives: direct object
(12a) LOB E01 83: We who love lace-craft hope that you will enjoy the work that this book offers ...

(12b) LOB E03 16: ... ensuring that the drill Ø you fancy will drive the attachments ...

The effect of this interactive variable is particularly strong in subject relatives, as can be seen from the table below, which shows the distribution in Quirk's data of relative markers in subject restrictive relatives with personal and non-personal antecedents. The distribution is very highly significant ($P < 0.001$).

Table 8 Relative markers in subject restrictive relatives (data from Quirk, 1968)

	wh	*that*	zero	Total
Personal	202 (91%)	19 (9%)	1 (.45%)	222
Non-personal	146 (48%)	157 (52%)	1 (.33%)	304
Total	348	176	2	526

$\chi^2 = 105.29$, $P < 0.001$ (Yates' correction applied)

To merge these two groups is to mix data from two distinct populations, and as a result Biber's analysis is further confounded. Note that the data can be classified manually (personal subject relatives, non-personal subject relatives; personal object relatives, etc), but the *automatic* identification and classification of antecedents is a difficult problem. First, to identify the antecedent of a relative clause involves the resolution of attachment ambiguities (Cardie, 1992), as illustrated in (13) below. Here the antecedent is not *business community* or *members of the business community*, but *attractions*.

(13) Brown H30 1770 ... but an educational institution can offer many potent intangible attractions to members of the business community that will offset the differences in income.

Secondly, the task of categorizing antecedents reliably can be difficult even for humans, and is not within the reach of current text analysis technology. Two examples will serve to make the point. The first shows an

antecedent whose referent must be determined from the discourse context, thus requiring some sort of reference resolution module (cf. Dahl and Ball, 1989). The second shows an animal antecedent, which is variably treated as personal or non-personal depending on the speaker/writer's point of view (Quirk *et al.*, 1985, p. 1245). Note that in (15c) we cannot tell how to classify the antecedent, since *that* is neutral.

(14) Referent not recoverable from immediate environment
1980 Pear Stories ES 10: A--nd one of ... they were probably the two there were see ... two that seemed about ... to be about his age ...

(15) Antecedent treated variably as personal or non-personal (or unclear)

(15a) 1986 Hearne Adam's Task 87: A dog who is track-sure is, most of all, undistractible.

(15b) LOB E32 93: ... I recently heard the scathing comment "slipped patella" used about the action of a dog which gave an occasional hop ...

(15c) 1989 Pictorial Encyclopedia of Dogs 122: Nicknamed 'the daredevil,' the Irish Terrier is an incredibly courageous dog that served as a sentinel and was used to carry messages on the battlefield during World War I and World War II.

In short, the analyst can control for variables known to affect the distribution of restrictive relative markers, but at the cost of manual effort. Still, a time-consuming but linguistically sound study is clearly to be preferred to an automated analysis with inaccurate results.

Summary

The investigation of variation in linguistic phenomena is a problem that calls for the use of large corpora, and genre-related variation is an important area of study. However, genre is not the only variable that must be controlled for: both linguistic and non-linguistic variables must be identified and taken into account. The research design that contains hidden variables is flawed, and no reliable conclusions can be drawn from the results. Unfortunately, the identification of known linguistic variables in a large corpus is time-consuming when done by hand, and in the case of factors affecting choice of relative markers it is seen that reliable automatic identification is not currently feasible. Large-scale macroscopic studies of multiple linguistic phenomena such as those conducted by Biber are thus premature. For the results to be reliable and obtained within a reasonable amount of time, such studies require first the support of adequate microscopic studies of individual phenomena, and secondly the development of robust text analysis tools or of large parsed corpora.

6. Conclusions

In this paper, we have reviewed a number of pitfalls of automated text analysis. Our goal is not to discourage the use and development of computational tools for analysing electronic text, but rather to suggest that they be used with a full awareness of their limitations, and that they be coupled with or replaced by manual methods when appropriate. The development of robust systems for processing unrestricted text is still a long-term goal, as Church (1991) points out and most computational linguists would probably agree: 'In the long term, at least, we have the responsibility to deliver a large grammar with broad coverage for unrestricted text. We need to start thinking now about how we could ever hope to achieve this goal.' (Church, 1991, p. 102)

In the meantime, we issue a call for serious consideration of the methodological issues in automated corpus analysis and the development of evaluation metrics for text analysis tools. Finally, we suggest that the automated analysis of large text corpora requires further work on the recall problem and on statistical techniques for estimating the reliability of results.

Acknowledgements

An earlier version of this paper was presented at the ALLC-ACH93 Conference at Georgetown University (June 1993). I gratefully acknowledge the assistance of members of the CORPORA list, in particular Knut Hofland and Lars Martin Fosse, who furnished me with several valuable references in connection with the LOB.

Notes

1. For a useful introduction, see Burnard 1992.
2. Examples from the LOB, London-Lund and Brown corpora in this paper are taken from the *ICAME Collection of English Language Corpora*, distributed on CD-ROM by the Norwegian Computing Centre for the Humanities.
3. Parsed corpora with zeroes identified include the Penn Parsed Corpus of Middle English (Taylor and Kroch, 1993).
4. As Levin (1987, p. 176) points out, however, not all theories of grammar treat gappings as clauses.
5. The Brown and LOB corpora are in fact accompanied by word counts for each text and text group. But the counting of words in spoken data is not so straightforward, it should be noted. Approximate figures for each sample in the London-Lund Corpus are given in the documentation, but to obtain precise counts, difficult decisions must be made about whether to include interjections, repetitions, partial words, and so on.
6. The notions *mean clause length* and *word/clause ratio* may not be equivalent, it should be noted. For example, in computing the word/clause ratio for a text, we divide total word count by number of clauses. But in computing mean clause length, we might wish to first identify each clause, and then count the number of words in it. In spoken data, the latter method would exclude a large amount of non-clausal material, including false starts and utterances such as *No, Yeah* and *Absolutely*.
7. Johansson and Hofland (1989, p. 17) report mean sentence lengths for LOB, but not mean clause lengths.
8. Variables known or believed to affect choice of relative markers include date, regional variety, social class, sex, text type, grammatical role, antecedent type (personal versus non-personal), type of antecedent (quantifiers, superlatives, *same, only*, demonstratives), length of relative clause, and whether the clause is extraposed or adjacent to the head. For further discussion, see Quirk (1957), Kikai *et al.* (1987), Ball (1992).

References

Ball, C. (1991). *The Historical Development of the It-Cleft*. Ph.D. dissertation, University of Pennsylvania.

—— (1992). A Diachronic Study of Relative Markers in Spoken and Written English. Paper presented as NWAVE 21, Ann Arbor, MI.

Beaman, K. (1984). Coordination and Subordination Revisited: Syntactic Complexity in Spoken and Written Narrative Discourse. In D. Tannen (ed.), *Coherence in Spoken and Written Discourse*. Ablex, Norwood, NJ, pp. 45–80.

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

—— (1988). *Variation across Speech and Writing*. Cambridge University Press, Cambridge.

—— (1990). Methodological Issues Regarding Corpus-based Analyses of Linguistic Variation. *Literary and Linguistic Computing*, 5: 257–69.

Biber, D. and Finegan, E. (1988). Historical Drift in Three English Genres. In T. Walsh (ed.), *Georgetown University Round Table on Language and Linquistics 1988*. Georgetown University Press, Washington, DC, pp. 22–36.

Blair, D. and Maron, E. (1985). An Evaluation of Retrieval Effectiveness for a Full-text Document Retrieval System. *Communications of the ACM*, 28.3: 289–99.

Burnard, L. (1992). Tools and Techniques for Computer-assisted Text Processing. In C. Butler (ed.), *Computers and Written Texts*. Basil Blackwell, Oxford, pp. 1–28.

Cardie, C. (1992). Corpus-based Acquisition of Relative Pronoun Disambiguation Heuristics, *Proceedings of the 30th Annual Meeting of the Association for Computational Linguistics*, pp. 216–23.

Chafe, W. (ed.) (1980). *The Pear Stories: Cognitive, Cultural and Linguistic Aspects of Narrative Production*. Ablex, Norwood, NJ.

Church, K. (1991). Review of Aarts and Meijs (eds) 1990, *Theory and Practice in Corpus Linguistics, Computational Linguistics*, 17: 99–103.

Collins, P. (1991). *Cleft and Pseudo-Cleft Constructions in English*. Routledge, London.

Crystal, D. and Davy, D. (1969). *Investigating English Style*. Indiana University Press, Bloomington.

Dahl, D. and Ball, C. (1989). Reference Resolution in PUNDIT. In P. Saint-Dizier and S. Szpakowicz (eds), *Logic and Logic Grammars for Language Processing*. Ellis Horwood, Chichester.

Francis, W. and Kucera, H. (1982). *Frequency Analysis of English Usage: Lexicon and Grammar*. Houghton Mifflin, Boston, MA.

Gibson, J., Gruner, C., Kibler, J., and Kelly, F. (1966). A Quantitative Examination of Differences and Similarities in Written and Spoken Messages, *Speech Monographs*, 33: 444–51.

Hearne, V. (1986). *Adam's Task*. Alfred A. Knopf, New York.

Johansson, S. and Hofland, K. (1989). *Frequency Analysis of English Vocabulary and Grammar Based on the LOB Corpus*. Clarendon Press, Oxford.

Kikai, A., Schleppegrell, M., and Tagliamonte, S. (1987). The Influence of Syntactic Position on Relativization Strategies. In K. Denning, S. Inkelas, F. McNair-Knox, and J. Rickford (eds), *Variation in Language: NWAV-XV at Stanford*, pp. 266–77.

Levin, N. (1987). A Pragmatic Concomitant of Gapping, *Proceedings of the Fourth Eastern States Conference on Linguistics*, pp. 176–86.

Long, C. and Strader, B. (1989). *The Pictorial Encyclopedia of Dogs*. Gallery Books, New York.

Nevalainen, T. and Raumolin-Brunberg, H. (1989). A Corpus of Early Modern Standard English in a Socio-historical Perspective, *Neuphilologische Mitteilungen* XC.

Norwegian Computing Centre for the Humanities (dist.) (1991). *The ICAME Collection of English Language Corpora on CD-ROM*.

O'Donnell, R. (1974). Syntactic Differences between Speech and Writing, *American Speech*, 49: 102–10.

Quirk, R. (1957). Relative Clauses in Educated Spoken English, *English Studies*, 38: 97–109. Reprinted in R. Quirk (1968), *Essays on the English Language, Medieval and Modern*. Indiana University Press, Bloomington.

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

Rissanen, M. (1991). On the History of *That*/Zero as Object Clause Links in English. In K. Aijmer and B. Altenberg (eds), *English Corpus Linguistics*. Longman, London, pp. 272–89.

Rydén, M. (1966). *Relative Constructions in Early Sixteenth Century English*. Almqvist and Wiksells, Uppsala.

Salton, G. (1989). *Automatic Text Processing*. Addison-Wesley, Reading, MA.

Sundheim, B. (ed.) (1991). *Proceedings of the Third Message Understanding Conference*. Morgan Kaufmann, San Mateo, CA.

Taylor, A. and Kroch, A. (1993). The Penn Parsed Corpus of Middle English: A Syntactically Annotated Database. Paper presented at the Pre-session on Corpus-Based Linguistics, Georgetown University Round Table on Languages and Linguistics, Washington DC.

Ward, G. (1988). *The Semantics and Pragmatics of Preposing*. Garland, New York.