ICAME conferences are invariably innovatory: new descriptions from new corpora using new techniques. In this regard, ICAME19-98 was no exception, but three new dimensions made it special: widespread use of very large corpora such as the *Cobuild Bank of English (BofE)*, the *British National Corpus (BNC)*, CD-ROM newspaper corpora, and even the internet (or worldwide web); as a result, the use of corpora for lexical studies; and the development of parsed corpus concordances.

No doubt the debate over the relative merits of corpus size versus register balance will continue. But there is certainly merit in pursuing size if, as these papers show, it yields good results. According to Kjellmer, for instance, ‘the arrival on the scene of large-scale language corpora has now made it possible for us to register even very marginal variation phenomena in a fairly systematic way, phenomena which may thus be the precursors of changes to come’ (this volume: 115).

The spoken component of the *BofE* (currently 325,000,000 words) is used in the paper by Susan Blackwell as a resource of data for discourse uses of three words: honest (on its own, but also its various collocations to be (quite) honest, let’s be honest, I’ll be (quite) honest, and honest to God), look (especially utterance-initially), and well when tagged as a ‘formulaic interactive expression’. A very different use of the *BofE* features in the theoretical paper by Oliver Mason on the possibilities and techniques for measuring collocations. Mason’s purpose is to demonstrate that ‘the span for computing a word’s collocates can be determined empirically’; to this end, he uses the *BofE* as a control corpus, and a sample of it for detailed investigation. The *BofE* is also used by Vincent Ooi to generate examples of words under investigation by him as Asian words. The occurrence of some ‘eastern’ words in this ‘western’ corpus may be explained by the *BofE*’s high newspaper constituency.

The *Cobuild Direct Corpus (CDC)* (50 million words of the *BofE* available as a CD-ROM) is used by Göran Kjellmer in his study of complements of the lexical verb TRY. He finds that, whereas the TRY + bare infinitive neither occurs in any of the earlier corpora (the *Brown Corpus*, the *Lancaster-Oslo/Bergen Corpus* and the *London-Lund Corpus (LLC)*, nor in the *Freiburg LOB* Corpus (the 1991 equivalent of *LOB*), there are 47 occurrences in *CDC*, with most examples coming from ‘Australian News’ and ‘UK Spoken’ sections. Interpreting the behaviour of these data using criteria for auxiliaryhood, Kjellmer concludes that TRY is beginning to move towards auxiliaryhood status.

The main alternative to the *BofE* is the 100,000,000 word *BNC*. The webreport by Lou Burnard http://users.ox.ac.uk/~lou/reports/9805icame.htm finds that nearly 20 of the 70 or so papers and posters at the conference made use of *BNC* data, but this has primarily to do with accessibility and cost, now that the
BNC is available over the internet (http://www.thetis.bl.uk). Although the BNC can only be exploited through its own dedicated retrieval software, SARA (cf. Aston and Burnard 1998), an attractive web-based browser with a user-friendly point-and-click interface, BNCweb, is described by Hans Martin Lehmann, Peter Schneider and Sebastian Hoffmann. In a separate paper, Sebastian Hoffmann and Hans-Martin Lehmann extracted from the BNC some 55 low frequency collocations (i.e. word combinations sequences of up to three words either side of the node occurring less than 100 times) with high log-likelihood values and then elicited familiarity of these collocations among native and non-native speakers. They found that native speakers knew, and could complete, far more collocations than non-native speakers. They also investigate a supposed connection between frequency and familiarity but are unable to confirm it.

Subsets of both spoken components of the BNC and the BofE are combined in Tony McEnery, Paul Baker and Andrew Hardie’s paper on very colloquial language, which reports their problem-oriented Lancaster Corpus of Abuse. This new corpus of lexically-oriented speaker-annotated citations is designed to facilitate investigation of a particular lexical class. It will likely reassess the existing scholarship and provide a fresh socio-pragmatic analysis of individual items as used in the 1990s.

Jürgen Gerner uses the spoken component of the BNC for an exhaustive investigation of the choice and grammatical number expressed in anaphoric pronouns of the indefinite personal pronouns: someone/somebody, anyone/anybody, everyone/everybody, and no-one/nobody. Its main finding is that plural anaphors is the norm – especially the use of they with singular reference. Gerner shows that use of singular they is not influenced by age or social class of the speaker, or by context or region.

Inge de Monnink searched the BNC for discontinuous modifiers as .. a, so ... a, and too ... a in a span of 5 words within noun phrases, specifying that as, so and too are adverbs (tagged AV0) and a is an article (tagged AT0). Most examples occurred in non-fictional written texts.

The spoken component of the BNC is receiving particular attention. Along with the LLC and the Reading Emotional Speech Corpus, it forms part of the CHRISTINE Corpus, which is being annotated by Anna Rahman and Geoffrey Sampson with a rigorously systematic part-of-speech classification scheme based on, but being developed from, the earlier SUSANNE scheme which had been applied to written texts. Their paper shows the many types of difficulties for tagging which are presented by spontaneous speech.

Part of the spoken component of the BNC comprises the Bergen Corpus of London Teenage Language (COLT) compiled by Anna-Brita Stenström. Two papers compare teenage language with adult language. Anna-Brita Stenström shows how well and enough are becoming refunctioned as adverbial intensifiers. Carita Paradis compares degree modifiers of adjectives with those occurring in the LLC and finds real evidence of change, which will have consequences for present-day lexicography.
Both the BNC and the CDC are used in a study by Aimo Seppänen and Joe Trotta of the complex but marginal wh- + that relative clause construction (e.g. *what a mine of useless information that I am!*), eliciting 37 and 53 examples respectively.

These papers confirm the value of very large corpora for descriptive studies. A further source of very large amounts of data are newspaper corpora which are available for purchase on CD-ROM. Although these usually prevent exploitation of grammatical items, they are useful for lexical or lexis-based syntactic investigations. An original champion of CD-ROM newspaper corpora is David Minugh, who uses CD-ROM corpora of the *Times* and the *Independent* from the UK, of the *New York Times* and the *Los Angeles Times* from the US, and the *Sydney Morning Herald* from Australia. He also uses the American Broadcast News transcripts to examine a subset of 200 idioms from the Collins Cobuild Dictionary of Idioms, demonstrating the wealth of material for idiom investigation both in corpus linguistics and in foreign language teaching. Some of these CD-ROM newspaper corpora (New York Times and the Independent) are used in the study by Hans Lindquist of inflectional and periphrastic comparative forms of some 90 disyllabic adjectives. In addition to the Bank of English, Vincent Ooi uses the *Straits Times* (Singapore’s English broadsheet), the *New Straits Times* (Malaysia’s English broadsheet) and the *New Paper* (Singapore’s English tabloid) to study words and especially collocations with a specifically Asian distribution or connotation. He concludes that different communities have their own Englishes, and that global dictionaries should record this information. Use of newspaper corpora by linguists seems certain to increase.

For all the scope and benefits demonstrated in these papers, the limits of the BNC are tested in the paper by Magnar Brekke on the basis of a lexical study of *chaos* and *quantum*. Whereas these two words hardly merit a paper to themselves, despite their 1,424,780 occurrences largely as part of complex multi-word technical terms, compared with 2,425 occurrences in the BNC, Magnar Brekke, following on from the syntactic paper by Bergh et al. (1998) presented at the 1997 ICAME Conference, shows convincingly that the entire internet lends itself for exploitation as a corpus of randomly-occurring data, with impressive results. There is no larger machine-readable corpus than the Cybercorpus.

All the same, corpora comprising pre-determined samples of a fixed size continue to be compiled. Perhaps the most important are the national components of the International Corpus of English, of which the British Component of the International Corpus of English (ICE-GB) is the only one so far to be tagged and parsed and available on CD-ROM. Sean Wallis, Bas Aarts and Gerald Nelson show how the project’s dedicated software, the International Corpus of English Corpus Utility Package (ICECUP) can be used to generate display outputs based on syntactic parses using its highly user-friendly ‘Fuzzy Tree Fragment’ search tool. Another two such are the Wellington Corpus of Written New Zealand English and the Wellington Corpus of Spoken New Zealand English, which are used here for a study by Graeme Kennedy and Shunji Yamazaki of words of Maori origin, which are also treated in the recently-
A small corpus of text samples of fiction, non-fiction, drama and spoken conversation forms the basis of a study of verb phrase structure by Nelleke Oostdijk in which detailed analysis is given to each of the verb phrases in a 19,368 word extract from a biological textbook.

Another type of corpus is the special or single register corpus. One such is the Louvain Corpus of Native English Essays (LOCNESS), which is used as a control corpus for the various national learner-essay components of International Corpus of Learner English (ICLE). Using type-token ratios, Sylviane Granger and Martin Wynne investigate the lexical density of essays produced by Dutch, French, Polish and Spanish native-speakers in comparison with those by English native-speakers.

The register of medical texts forms the basis of two studies. A 400,000 word corpus of contemporary American medical texts is used for a register study by Minna Vihla of epistemic possibility. The Corpus of Early English Medical Writing 1375-1750, is used for a study by Irma Taavitsainen of personal and impersonal metatextual and metadiscursive comments of the type it is to be said, or as I said or as thou shalt understand. Since the renaming of ICAME as the International Computer Archive of Modern and Medieval English, it is particularly pleasing that a paper based on a historical corpus is included in the conference proceedings volume.

A fiction corpus comprising the novels of Robert Louis Stevenson is used by Pasi Tapanainen and Timo Järvinen in the development of the concordancers which operate on tagged texts for the purpose of syntactic analysis. The paper also shows how, on the basis of of zero-relative pronouns, elided items can be concordanced. A second paper, by Atro Voutilainen, which builds on the English Constraint Grammar (EngCG) tagger-parser, shows display outputs of the revised EngCG-2 tagger-parser, the new EngLite tagger-parser (with only 12 tags), and the new Functional Dependency Grammar (FDG) parser, with its visual tree diagrams, all developed in Helsinki, is based on an ad hoc corpus of only 2,332 words comprising three short excerpts from a newspaper, a travel diary and a scientific text. As corpora grow larger, so they also become smaller!

One of the smallest must be the collection of transcripts of spoken police interviews defendants, and of textual documents which are alleged to have been made by defendants, but which the defendants deny having written, which is reported on by Susan Blackwell. In this forensic case study, for which the BofE is used as a control corpus, Susan Blackwell is able to reach conclusions about the authenticity and authorship of the disputed written statements.

With the abundance and diversity of machine-readable corpora upon which the present studies are based, the volume will certainly live up to its name of Corpora Galore. Many other paths through these papers could, of course, be taken.

The papers are grouped into three sections. The first comprises lexical studies, which are now possible because only very large corpora such as newspaper corpora or the BNC yield sufficient tokens of any item as evidence for
its semantic or collocational behaviour. The lexical papers discuss the following topics: frequencies of *look* and *honest* as discourse markers (Susan Blackwell); frequencies of collocations such as *false pretences* and *prohibitively expensive* as well as their use by native and non-native speakers (Sebastian Hoffmann and Hans Martin Lehmann); Maori words in New Zealand English such as *kiwis* and *pakehas* (Graeme Kennedy and Shunji Yamazaki); swearing and abusive words such as *prick*, *cunt*, and *moron* and the sex of their addressee (Anthony McEnery, Paul Baker and Andrew Hardie); frequencies of idioms such as *cutting edge* and *flesh and blood* (David Minugh); and specifically ‘eastern’ or Asian collocations of words such as *filial piety* or neighbourhood *school* (Vincent Ooi).

The syntactic and semantic descriptive topics dealt with in the second section are: the pronominal forms which refer back to indefinite pronouns as in *everybody ought to do what (he/she/they) can or this is what happens when someone leaves a job of (his/her/their) own accord* and, in particular, the use of singular *they* (Jürgen Gerner); the apparently increasing auxiliarisation of the lexical verb *try* followed by a bare infinitive as in *... meetings to try clarify its tactics ... or try not give them anything to eat* (Göran Kjellmer); inflectional and periphrastic forms of disyllabic comparative and superlative adjectives as in *deadlier* and *more deadly, deadliest and most deadly* (Hans Lindquist); discontinuous modifiers in the noun phrase as in *a rumour was spread that the president was dead* or noun phrase modifiers which have moved their position as in *I did none too good a job as it turned out* (Inge de Monnink); adverbal degree modifiers such as *frightfully good* and *well nice* and their semantic interpretation (Carita Paradis); the double relative construction involving a *wh*-expression as well as that as in *it’s surprising how much stuff that collects under here or I yielded to whatever arguments that were given or I don’t know why that you go for a certain colour* (Aimo Seppänen and Joe Trotta); adverbial intensifiers such as *it’s enough funny* or *that’s well nice* occurring before simple adjectives rather than participial adjectives (e.g. *well developed*) (Anna-Brita Stenström); phrases used to organise texts by way of authorial comment (Irma Taavitsainen); and the expression of epistemic possibility through modal auxiliary verbs (can, could, may, might) and modally harmonic adverbial expressions (maybe, perhaps, possible, possibly and possibility) (Minna Vihla).

In the past, ICAME papers have not only been descriptive, they have been concerned with theoretical issues, first principles, and methodological innovation. The third section of papers deals with the following topics: the use of the internet as a cybercorpus (Magnar Brekke); use of type-token lexical ratios as an index of learner ability as identified by lexical over- and under-differentiation (Sylviane Granger and Martin Wynne); a user-friendly point-and-click web based client or browser for accessing the *BNC* (Hans Martin Lehmann, Peter Schneider and Sebastian Hoffmann); frequencies, type-token ratios and entropy values of the preposition of (as in *the best X of Y*) co-occurring with the certain as evidence that collocation is bound up with colligation (Oliver Mason); simplicity and complexity of verb phrase structure to show the relationship between linguistic delicacy and annotation schemes for tagging and
parsing (Nelleke Oostdijk); tagging schemes for spontaneous speech (Anna Rahman and Geoffrey Sampson); searching and concordancing tagged and parsed corpora and displaying the output (Pasi Tapanainen and Timo Järvinen; Atro Voutilainen; Sean Wallis, Bas Aarts and Gerald Nelson).

Through these 24 papers, the volume presents analyses and techniques in describing English, as announced by its subtitle. They were chosen because their descriptive findings amounted to potentially lasting contributions to their self-contained topics, or because the innovatory techniques would be of benefit to other corpus linguists both within the ICAME community and more generally, and that it was important for the information to be published promptly. Increasingly, references are to websites, which readers are encouraged to consult, not least for any update on the corpus material or exploitative techniques presented here.

References
