Reviews

Clive Souter and **Eric Atwell** (eds.). *Corpus-based computational linguistics*. Amsterdam, Atlanta, Ga.: Rodopi, 1993. 260 pp. ISBN 90-5183-485-3. Reviewed by **Christian Mair**, Albert-Ludwigs-Universität Freiburg, Germany.

The good ship 'ICAME' is sailing on. Not only are the proceedings of the annual conferences usually published; but there have recently also been a number of *festschrifts* and other collections devoted to corpus linguistics in general and work by ICAME regulars in particular. So the body of literature which the present volume – essentially the proceedings of the twelfth ICAME conference staged at Ilkley/Leeds in 1991 – has been added to has grown to considerable proportions.

Over the years ICAME meetings have provided a forum for diverse and not always easily compatible interests. This is reflected in the tripartite division of the present volume. Part 1, 'Corpus collection, annotation and standards,' with contributions by Bauer, Greenbaum, Meyer/Tenney, Burnard, Knowles and Piepenbrock, deals with the progress made in individual corpus projects or with general issues arising in the compilation of corpora and language databases. Part 2, 'Corpus applications in linguistics,' represents the philological side of the field, with papers by Wichmann, Schmied, (Christine) Johansson, Noel, Wikberg and de Haan exploring fine points in English structure and usage with the help of available corpora. Part 3, 'Corpus-based syntactic and semantic analysis software,' with contributions from Gorman/Hardy, Souter, Atwell, Wilson/Rayson, Guthrie, Cowie and Meijs, discusses advances in tagging and parsing.

In a short review it is obviously impossible to discuss the individual contributions in detail. This is bad but not entirely disastrous as many of them are progress reports on projects already familiar from previous ICAME proceedings. In what follows, I shall essentially provide some selective comment where I consider it useful.

In terms of novelty, the highlight of Part 1 is Laurie Bauer's introduction to the long-awaited Wellington corpus, the New Zealand Brown/LOB clone. Discussing a selection of usage problems, Bauer shows that our knowledge of New Zealand English norms is minimal and likely to be extended greatly through corpus-based research, but at the same time the size of one million words forces the analyst into inspired guesswork for lack of a sufficient amount of relevant data[1]. The remaining papers report on advances in the ICE, Lancaster/IBM and CELEX projects and the Text Encoding Initiative. ‘Tagger’, an interactive tagging program, was developed for use in the American ICE component but has the potential for wider application.

The descriptive section covers topics in phonetics/intonation studies, syntax (two papers on aspects of relative clauses, that perennial favourite of ICAME grammarians, one on noun complementation), and stylistics. In their separate ways, all these contributions allow the reader glimpses into the corpus linguist’s workshop and toolshed, and as is to be expected, they are very solid on frequencies and distributional patterns, while the heretic with a more speculative and theoretical orientation may well occasionally be asking whether she/he really wanted to find out about what is presented. Verbs in the ‘Love and Romance’ Section of the Brown Corpus, verbs in written American English, verbs in English, or verbs – what is the right level of abstraction, and how much theory is necessary for the honest corpus-linguistic craftsman?

The third section is somewhat more voluminous than usual in ICAME publications. As an outsider to the tagging and parsing field, one admires the ingenuity and expertise that goes into the solution of numerous analytical problems but in the end one cannot help wondering whether the computational analysis and modelling of natural languages is not a goal that is beyond our reach for reasons of principle. This is not to be understood as a call for defeatism as especially in this field even minimal advances may have beneficial effects in numerous linguistic fields. It is interesting to note that many papers in this section focus on aspects of the lexicon, and the harmonising of lexicon and grammar, thus paralleling a development in general linguistics, where such a shift of attention away from formal syntax has also been in evidence for some time.

But back from the outer reaches of the philosophical debate on the possibility of ‘artificial’ intelligence to the immediate concerns of this review. *Corpus-based computational linguistics* is an interesting and generally well-edited collection of essays and a welcome addition to the existing series of ICAME conference proceedings.

Note

1. On p. 8, the misleading ‘irregular plural’ should be changed into ‘irregular past’.

Charles Meyer. *Apposition in contemporary English.* Cambridge: Cambridge University Press, 1992. 152 pp. ISBN 0-521-39475-9. Reviewed by **Nelleke Oostdijk**, University of Nijmegen.

In the tradition of descriptive grammar apposition in contemporary English has been studied extensively. However, there appears to be little agreement as to how apposition should be defined. A survey of past treatments of apposition leads Meyer to conclude that they 'provide either an inadequate or an incomplete definition of apposition' (p.3). While apposition has been discussed in most grammars, Meyer's book is the first full-length treatment of the subject.

In his preface Meyer observes that his intention with this study is two-fold: he intends not only to define apposition, but also to detail its usage in computer corpora of spoken and written British and American English, in an attempt 'to clarify the confusion surrounding the category of apposition' (p. xiii). Unfortunately, what follows does not live up to the expectations raised. While the reader is repeatedly frustrated by the fact that Meyer does not explicitly define his descriptive notions – so that one can only make an educated guess at what they denote, thereby making it impossible to rate his findings at their true value – Meyer in general fails to convince where the linguistic argumentation is concerned. Another general point of criticism concerns the nature and presentation of his statistical analyses of his data. Below I shall illustrate these points of criticism while making an attempt at discussing the contents of the book chapter by chapter.

Meyer's perception of apposition as 'a grammatical relation ... realized by constructions having particular syntactic, semantic and pragmatic characteristics' (p. xiii) is reflected in the structure of the book. Apart from the preface and two appendices (one listing the grammatical tags used in the problem-oriented tagging, the other listing the number of appositions found in the individual samples of the corpora), there are five chapters, each of which focuses on a particular aspect: 1. Apposition as a grammatical relation; 2. The syntax of apposition; 3. The semantics of apposition; 4. The pragmatics of apposition; 5. Apposition in the grammar of English.

1. *Apposition as a grammatical relation*

The first section of this introductory chapter is devoted to a cursory discussion of previous studies of apposition. Without exception Meyer finds these to be inadequate and/or incomplete. He finds it unsatisfactory to restrict apposition, as some studies do, to two juxtaposed noun phrases, claiming that it is arbitrary and ignores obvious linguistic similarity. On the other hand, he adds, expanding the class of apposition too much renders the notion meaningless. While from previous studies possible criteria have emerged for identifying appositions (e.g. juxtaposition of the appositive units, separability by means of a marker of apposition), there is as yet no comprehensive linguistic description which encompasses a discussion of the syntactic, semantic, and pragmatic characteristics of apposition. Against this background Meyer (p. 5-6) states that

> '... it is providing this kind of comprehensive linguistic description of apposition that is the goal of this study. Apposition, it will be demonstrated, is best viewed as a grammatical relation that stands in opposition to relations such as complementation or modification. ... Defining apposition in the manner proposed in this study avoids the inadequacies of past treatments of apposition. Viewing apposition as a grammatical relation having various realizations does not arbitrarily restrict the class of apposition to only certain kinds of constructions. At the same time, the class of apposition is expanded in a principled manner so that only certain kinds of constructions are considered appositional. Finally, the linguistic characteristics posited to define apposition cover not just some characteristics but all of them.'

The reader may find it unfortunate that Meyer's definition of apposition as a grammatical relation remains implicit, as he merely describes the linguistic characteristics of units in apposition, without specifying what the relationship amounts to. In this respect, the arguments he brings to bear against previous treatments of appositions contribute to confusing the reader rather than succeeding in clarifying things. For example, in his discussion of Matthews (1981), who defines apposition as a type of juxtaposition, Meyer observes that 'strictly speaking, apposition is not a type of juxtaposition, since it is possible for many units in apposition not to be juxtaposed' (p. 5). The counterexample that Meyer gives at this point is the following:

> *Three people* attended the meeting: *Dr. Smith, Professor Jones, and Mr. King.*

It seems to me that Meyer confuses the grammatical relation of 'juxtaposition' with the linear notion of 'adjacency': while the noun phrases can be said to be juxtaposed, they are not adjacent. Additional arguments provided by Meyer against Matthews' view appear unfounded. For instance, when Meyer observes that 'because apposition is a grammatical category that is realized by so many different kinds of constructions, it makes more sense to say that apposition is a relation itself rather than an instance of another type of relation, juxtaposition' (p. 5) this hardly qualifies as a sound linguistic argumentation.

The two sections that follow contain brief descriptions of the computer corpora which Meyer used to investigate apposition and of his method of analysis. Three corpora were involved: the London-Lund Corpus (LLC), the Survey of English Usage Corpus (SEU) and the Brown Corpus (Brown). Approximately 120,000 words of each of the corpora were investigated. The material was selected so as to enable comparisons between different varieties of English: (1) spoken vs written English; (2) American vs British English; (3) varieties within the spoken genre; and (4) varieties within the written genre.

While the title of the first chapter is not very apt (only the first section has some bearing on the subject-matter referred to in this title), the heading of the final section of this chapter, viz. 'The computational analysis of appositions in the corpora', is entirely out of place. In this section Meyer gives a description of his method. Basically, it amounts to a procedure of problem-oriented tagging – by hand, as it turns out – after which the data were made accessible by means of a spreadsheet, while a statistical package (SPSS) was used to yield frequency counts. Two points of criticism can be brought to bear at this point. The first concerns the problem-oriented tagging. Instead of outlining to the reader the linguistic motivation for the tags that were used in the tagging procedure and detailing how they were applied, Meyer refers to Appendix 1 which merely lists the tags, without any comment whatsoever. The reader who sets his hopes upon what is to follow in any of the subsequent chapters is disappointed: while a number of tags get explained along the way, this certainly does not hold true for all the tags that are listed and their application remains obscure. The second point of criticism concerns Meyer's use of SPSS. It is a pity he has not used the full resources of SPSS: this might have provided us with some statistically well-founded argumentation, which is now absent in the book (Meyer only presents fairly straightforward counts).

2. The syntax of apposition

In the second chapter Meyer discusses the syntactic characteristics of apposition. He argues that apposition is a gradable grammatical relation in the sense that 'some appositions are fully appositional' while 'other appositions behave in a manner that places them on a gradient between apposition and other grammatical relations' (p. 10). A set of criteria is postulated that serves to determine the amount of interdependency between the units of an apposition and thus the question whether an apposition must be considered to be central or peripheral. The criteria are the following:

(1) the first unit of the apposition can be optionally deleted;
(2) the second unit of the apposition can be optionally deleted;
(3) the units of the apposition can be interchanged.

A central apposition, according to Meyer, is one to which all criteria apply (so that the units in apposition can be said to be structurally independent), while with syntactically more peripheral appositions this is not the case. Although the idea of viewing apposition thus as a gradable relation undoubtedly has its merits, it is not entirely clear how the set of criteria must be applied: all criteria seem to be equally important, and therefore an apposition which fails to satisfy one or more criteria automatically qualifies as peripheral. However, this seems to be contradicted by Meyer's observations: while in first instance (p. 41) he posits that appositions 'that are coordinative will be considered central appositions', later on (p. 44) he observes that 'there also exist instances of coordinative apposition in which the units cannot be reversed.'

While apparently Meyer takes the extent to which the units in apposition are structurally independent to be the most important syntactic characteristic of apposition, he also discusses the syntactic form of units in apposition and their syntactic function, as well as what he refers to as the linear and hierarchical structure of apposition. We can only assume that for the description of the syntactic form and function of appositions Meyer adopts the descriptive model one finds in Quirk *et al.* (1985). The fact that the definitions of descriptive notions remain implicit seriously hinders the interpretation of the results and at times makes it impossible to follow Meyer's argumentation. By way of illustration, let us take a look at some tables in this chapter.

At the beginning of the chapter, a table (2.1) is presented which gives an overview of the distribution of the syntactic forms of the units in apposition that were found in the three corpora. In the course of the chapter there is a gradual breakdown of these overall figures. In Table 2.1 Meyer distinguishes four general types: (1) nominal apposition; (2) NPs in apposition with clauses or sentences; (3) apposition with obligatory markers of apposition (NP + NP; Other); (4) non-nominal apposition. Nominal appositions consist of units that are 'various kinds of noun phrases': proper NP, common NP, miscellaneous NP and pronoun (cf. Table 2.2). Surely a reader would like to know what constitutes a 'proper NP', especially if in the classification in Table 2.3 proper NPs are distinguished from NPs such as 'NP lacking determiner', 'definite NP', etc., and the notion 'proper NP' appears to be interchangeable with 'proper noun'. One can only speculate as to how Meyer would classify an NP like 'this Mr Jones who came to see you last week'.

Questions are also raised by the fact that in the text Meyer observes that 'the appositions in the corpora, as Table 2.1 shows, consisted of two units that had four general forms' (p. 10). The reader will find himself puzzled by this observation: while Meyer here classifies all appositions that were found in the corpora as appositions that consisted of two units, this is repeatedly contradicted in subsequent sections and chapters. For example, in chapter 5 Meyer summarizes his findings as follows: 'Even though more than two units can be in apposition ..., most units (92 percent) were single appositions consisting of only two units ...' (p. 124). The class of nominal appositions is also problematic in so far as the delimitation of the subclasses is concerned. One subclass is that of 'proper NP'. According to Meyer's definition this subclass consists of appositions in which one or both units are proper nouns (i.e. one or both units are NPs headed by a proper noun?). At the same time, however, he defines a class of appositions in which one or both units are pronouns (i.e. NPs headed by a pronoun?). Consequently, appositions in which one unit is a proper noun, while the other is a pronoun (to use Meyer's terminology) can be classified either as a proper NP or as a pronoun. Seen in this light, Meyer's observation that the one type of apposition occurred much more frequently than the other remains unsubstantiated: the difference in the frequency of occurrence must possibly be ascribed to inconsistency in the classification.

With regard to the syntactic function of appositions Meyer observes that 'appositions in the corpora typically had functions associated with noun phrases' and a fairly large number of the appositions had functions 'associated with positions ... that promoted the principle of end-weight' (p. 35). In his problem-oriented tagging Meyer distinguishes nine functions: non-existential and existential subjects, direct, indirect and prepositional object, subject and object complement, adverbial and verb. In so far as 'a syntactic function could be determined for at least one of the units in apposition' the apposition was tagged for one of these nine functions; when this was impossible, the tag 'no function' was applied. What may worry the reader is that the total number of appositions that occur in Table 2.26 (which gives an overview of the syntactic functions of units in apposition) is only 2,407, while the total number of appositions Meyer found in his corpora is 2,841. Are we to conclude that there are as many as 334 appositions to which the tag 'no function' applied? With regard to the distribution of appositions in this table Meyer posits that functions such as direct object, subject complement, prepositional object and subject of a sentence containing existential *there* promote end-weight and concludes that the high proportion (65%) of appositions that were found in these functions can be explained by the fact that appositions are heavy constructions. It seems to me that Meyer is somewhat hasty in drawing this conclusion: while generally speaking it may be true that certain functions promote end-weight, it is also true that whether a function such as direct object actually promotes end-weight or not depends on the position that this function takes in the sentence. Preposed direct objects do not carry end-weight. It also would have been more satisfactory if Meyer had been explicit about his approach to sentences with extraposed subjects and direct objects, cleft sentences, dislocations, etc. This would have given the reader a better idea of what exactly the various function classes encompass.

3. The semantics of apposition

While traditionally the relation of apposition is described as a relation consisting of two co-referential units, Meyer argues that adhering to this criterion would 'severely limit the constructions that can be admitted as appositions' (p. 57). He expands the number of semantic relations 'in order to admit as appositions a variety of different constructions'. Apart from co-reference, units in appositions can also be related, according to Meyer, by part/whole reference or cataphoric reference, or by a non-referential relation viz. synonymy, attribution or hyponymy. Among the examples that Meyer gives to illustrate each of these relations are the following:

> Linda dragooned *her uncle, Donald Murkland*, into a lunch the next day to find out what had happened. (co-reference)

> There were plenty of aristocrats, even in the great General staff, but there were *plenty of people* like *Ludendorff* who had absolutely no kind of family or anything. (part/whole reference)
>
> *This* is what Rivens wanted: *to introduce course one into Redford.* (cataphoric reference)
>
> The Evening News was *finished, consumed.* (synonymy)
>
> The jail authorities – attaching no particular significance to the episode – offered Barco whisky to revive him; but *the old fellow, a lifelong teetotaler*, refused it, and no more was thought of the matter. (attribution)
>
> The nitrogen in *organic matter (dead roots and shoots, manure, soil humus, etc.)* is changed during decomposition to an ammonium form ... (hyponymy)

In his discussion of the relations that were found to hold between the units of the appositions under investigation Meyer occasionally slips up. This is, at least in part, (again) due to his failure to clearly define the (syntactic) notions he uses. For example, in the table listing the forms of absolute synonyms (table 3.4) 58% of the cases are NP + NP appositions, while only 19% are PP + PP appositions. This leads Meyer to observe that ‘absolute synonymy, as Table 3.4 indicates, was most frequently a relation between units that were noun phrases, and less commonly a relation between units of other form classes’ (p.65). It seems to me that this conclusion depends largely upon the reliability of the classification of appositions as NP + NP or PP + PP. Going by the examples Meyer provides there is a great deal of obscurity surrounding the classification of appositions the units of which are noun phrases that follow prepositions: they can either be classified as PP + PP appositions, or they can be classified as unjuxtaposed (= non-adjacent) NP + NP appositions. Consider the example Meyer gives (p. 66):

> In order to present a concept as an object we should have to introduce the concept by means of a substantival expression; but Frege wishes to think of a concept as essentially something that can be represented *only* [italics in original] by *a nonsubstantival expression,* by *an expression that introduces its term in the verb-like, coupling propositional style.*

Where Meyer observes that ‘in appositions whose units were synonymous noun phrases, the noun phrases were indefinite or generic and occurred in two contexts: as juxtaposed noun phrases ... or (less commonly) as noun phrases functioning as objects of prepositions within prepositional phrases in apposition’ (p. 66) this observation is inherently contradictory: either we have apposition of noun phrases or the apposition is one of prepositional phrases, we cannot have appositions whose units are synonymous NPs while these NPs function as the prepositional objects within PPs in apposition.

The semantic classes that Meyer uses to describe the appositions are largely based upon the semantic classes in Quirk *et al.* (1985). He extends and redefines the classification given in Quirk *et al.* It is somewhat difficult to determine what exactly the consequences of these adjustments are. Quirk *et al.* (1985: 1308-1316) distinguish between strict restrictive and strict nonrestrictive apposition. For the latter type of apposition they introduce a semantic scale which runs from ‘equivalence (*ie* ‘most appositive’) to loose and unequal relationship (‘least appositive’), such as exemplification. It is only with strict restrictive appositions that they speak of the relative specificity of the units in apposition. Meyer seems to merge the two aspects, with the result that classes like ‘particularization’ and ‘exemplification’ which in Quirk *et al.*’s approach are described as ‘least appositive’, are now classified in the (superordinate) semantic class ‘more specific’, together with the classes of ‘appellation’ and ‘identification’ which in Quirk *et al.* are identified as ‘most appositive’. At the end of the chapter Meyer presents his semantic gradient of apposition. By this gradient those appositions are most appositive between whose units a relation of co-reference or synonymy holds, while appositions with units between which a part/whole relation exists are least appositive. It remains unclear where exactly in this description the various semantic classes fit in. The same goes for the notions ‘restrictive’ and ‘non-restrictive’ which Meyer has also discussed in the course of the chapter.

4. The pragmatics of apposition

The fourth chapter concentrates on the pragmatic aspects of apposition. Meyer discusses the ways in which pragmatic considerations influence the frequency and distribution of appositions. He argues that since ‘apposition is a relation in which the second unit of the apposition either wholly or partially provides new information about the first unit ... they are better suited to some contexts than to others and were therefore distributed differently across the genres of the corpora' (p. 92). While communicative factors account for the differences between speech and writing (in speech there is a tendency to have more 'old' information in the second units of appositions), other factors play a role as well. Meyer illustrates for example the stylistic function old information may serve.

In so far as markers of apposition are motivated by pragmatic considerations, they are discussed here. Meyer considers markers that are introduced into appositions for pragmatic reasons to be 'optional markers'. Consider the following example:

> It was shown that correction for secondary extinction was only necessary for *intense reflections,* namely *(310) and (400),* measured with the cO axis vertical.

Markers that have to be introduced for syntactic or semantic reasons are considered to be obligatory. In the corpora that Meyer investigated the optional markers were very uncommon: only 3% of the appositions contained an optional marker. The listing of the number of instances that have no marker <u>or</u> an obligatory marker under the header 'none' in the Table (4.3) is unfortunate. Moreover, I am sure that the reader is not so much interested in being informed of the fact that 97% had either an obligatory marker or no marker at all. It appeared that 84.3% had no marker, while 13% had an obligatory marker. Meyer observes that 'so few optional markers of apposition occurred in the corpora that it is difficult to determine precisely why they were so infrequent' (p. 98); rather than speculate at this point about the possible stylistic markedness of these markers, he could have investigated in how many instances of unmarked apposition there was potential for a marker. With the availability of this information a less speculative explanation might have been given.

5. Apposition in the grammar of English

In the final chapter of the book Meyer first summarizes the main syntactic and semantic characteristics of the appositions he found in his corpora: 'Syntactically, apposition is most commonly a relation between two juxtaposed noun phrases having a syntactic function (such as direct object) promoting end-weight' (p. 123) and 'semantically, appositions typically contain constructions that are referentially related and comprised of a second unit that adds greater specificity to the interpretation of the first unit' (p. 124). Then he discusses the variation in the distribution of appositions according to dialect and genre. On the basis of his findings Meyer concludes that 'variation in apposition usage is motivated not by differences between American and British English but by the varying functional needs of the different genres of English' (p. 126) There is, however, sufficient evidence in Meyer's data to posit that there *are* differences between American and British English. If only Meyer had looked (as I have done, cf. Table 1 below) at a cross-tabulation of the frequencies of apposition in the various genres in the two corpora (Brown and Survey), he would have found that while there are hardly any differences between the *overall* frequencies of appositions in the written genre for American and British English (49.7% vs 50.3%), and the variation by genre is considerable (21.6%, 36.5% and 41.9% for fiction, learned and press respectively), British English uses significantly more appositions in the learned genre than American English does (42.6% and 30.4% respectively), while in the category 'press' British English uses fewer appositions than American English (38.0% and 45.8% respectively).

Table 1

		Brown	Survey	Total
Fiction	abs. freq.	244	201	445
	rel. freq. (row)	54.8 %	45.2 %	100 %
	rel. freq. (column)	23.8 %	19.4 %	21.6 %
Learned	abs. freq.	312	442	754
	rel. freq. (row)	41.4 %	58.6%	100%
	rel. freq. (column)	30.4 %	42.6 %	36.5%
Press	abs. freq.	470	394	864
	rel. freq. (row)	54.4 %	45.6 %	100 %
	rel. freq. (column)	45.8%	38.0 %	41.9 %
Total	abs. freq.	1026	1037	2063
	rel. freq. (row)	49.7%	50.3%	100%
	rel. freq. (column)	100%	100%	100%

Meyer rounds things up by presenting his 'gradient of central apposition to peripheral apposition'. Syntactic and semantic characteristics, i.e. the degree of syntactic interdependence of the units in apposition and their semantic similarity, determine whether a construction is a central or peripheral apposition: most central are those appositions the units of which are syntactically independent and semantically co-referential, while the more structurally dependent the units are and the less semantically similar, the more peripheral the apposition.

All in all, in the light of the criticisms I have made I find it impossible to judge what insights Meyer has provided us with. Apposition remains a complex notion which I doubt will be understood in full before long.

References

Matthews, P.H. 1981. *Syntax*. Cambridge: Cambridge University Press.

Quirk, R., S. Greenbaum, G. Leech, and J. Svartvik, 1985. *A comprehensive grammar of the English language*. London: Longman.

Jan Aarts, **Pieter de Haan**, and **Nelleke Oostdijk** (eds.). *English language corpora: Design, analysis and exploitation*. Papers from the Thirteenth International Conference on English Language Research on Computerized Corpora, Nijmegen 1992. Amsterdam - Atlanta, GA: Rodopi, 1993. 312 pp. ISBN 90-5183-517-5. Reviewed by **Jürgen Esser**, Technical University of Aachen.

The 21 papers collected in this volume are presented in three sections dealing with the design and compilation of corpora, their grammatical analysis, and their exploitation. The first three articles deal with new historical corpora. Merja Kytö writes about the design features of the Corpus of Early American English and offers illustrative sample texts. Ian Lancashire reports on the project of a computer text-database of Renaissance bilingual and English-only dictionaries and glossaries. Computers make it possible to extract lexicographic information about Early Modern English from bilingual dictionaries, which existed before monolingual English dictionaries. (Shakespeare wrote without a dictionary.) This will shed new lexicographical light on Shakespeare's work. Susan Wright introduces the Cambridge Corpus of Early Modern English. In contradistinction to the set-up of the Helsinki Corpus, which is genre-oriented, this new corpus is author-oriented. It will enable scholars to study the development of personal styles. Wright for example shows how Addison changed his use of relative pronouns over 11 years.

The next three papers deal with new specialised corpora. Milena Collot and Nancy Belmore have compiled an Electronic Language Corpus. The language material is taken from nine different conferences of an electronic bulletin board system and comprises 200,000 words. With this kind of communication there is an addressor, an addressee and an audience. The corpus was tagged and analysed according to Biber's multidimensional-multi-feature model and compared with his six textual dimensions for a stylistic analysis. Sylviane Granger introduces the International Corpus of Learner English, a computerised corpus of essays written by advanced learners of English as a foreign language from various language backgrounds. And Cheng-yu Fang reports on a 1 million word corpus of the English of computer science to be established in Hong Kong.

The last three papers of the first section deal with the management of corpora. Gavin Burnage and Dominic Dunlop describe the standards and formats to be observed in the compilation of the British National Corpus which is to comprise 100 million words. Susan Blackwell addresses problems related to the use of newspapers in electronic form, e.g. *The Times*. There are external factors responsible for 'dirty data'. For example, words broken up into columns in the paper are not restored so that hyphens stay. And there are internal factors such as typos, deliberately employed non-standard spellings, older forms etc. The filter AVIATOR checks the vocabulary of the newspaper against a master list and the researcher has to decide whether a new form is to be included or not. Gerry Knowles desribes problems and procedures in the conversion of the Spoken English Corpus into a (vertical) database. The original prosodic transcription of orthographic texts is to be supplemented by a phonemic transcription taken from the computer-readable version of the *Oxford Advanced Learner's Dictionary*. Weak forms are generated by using information from grammatical tags (function words) and the prosodic transcription. (This seems to be a strange procedure for a corpus linguist, namely to predict and not record data.) Differences between textually conditioned word stress and dictionary word stress can be noted but no divergence in the vowel representations.

The mushrooming new corpora are certainly justified, but it must be irritating for the non-specialist to keep in touch with the development.

One of the functions of the ‘old’ Brown and LOB corpora, standardisation, is getting lost. Therefore publications like the one under review are much appreciated because they keep a wider audience informed.

The second section of the book is devoted to analytic procedures. Elizabeth Eyes and Geoffrey Leech, in their paper, focus on two areas: a description of the five projects under UCREL at Lancaster and a detailed report on the advancement in corpus annotation methodology. I found this very illustrative and suitable for the non-specialist. Hans van Halteren and Nelleke Oostdijk offer a glimpse into the workshop of tagging and parsing the TOSCA corpus at Nijmegen. One of the results is, for example, that the parser is much more successful in parsing fiction than it is in parsing non-fiction. As with the UCREL project, the interaction of human analyst and automatic computer procedures receives special attention. The next two articles stand out because they are highly technical and more in the area of computational linguistics. Mark-Jan Nederhof and Kees Koster report on their customised grammar workbench, which helps to construct large (modular) formal grammars. Ted Briscoe and Nick Waegner offer a solution to the problem of undergeneration, the situation that naturally occurring sentences are not correctly analysed. One of the computational problems is that with growing corpora size the search spaces (and processing time) have to be reduced drastically by grammatical constraints, higher initial probabilities and zero probabilities. Clive Souter reviews the different formatting styles used in the annotation of existing parsed corpora of English. His investigation is useful both for corpus linguists looking for standards of their own corpora and possibilities of importing data and also for the non-specialist, who gets a good impression of the discussed corpora, of which there are sample fragments given. The last section of the book deals with various aspects of corpus exploitation. In a sense this should be the section that offers the grapes of new linguistic insight. But due to the present state of the art and the nature of proceedings contributions, the methodological issues of corpus compilation and analysis outweigh the linguistic returns. This imbalance is even greater than it appears at first sight because the papers by Quinn and Collier belong more fittingly into the previous section on analysis. Akiva Quinn explains how object-oriented programming (as e.g. in MS Windows) is used for ICECUP, the utility program to be designed for the International Corpus of English. There is a clear introduction to the principles of object orientation and their implementation in utility functions such as searching, producing key-word-in-context concordances and statistics. The non-specialist reader gets a glimpse of ICECUP and acknowledges gladly that it is designed for linguists without any specialised computer knowledge due to the MS Windows standards. Alex Collier gives a clear introduction to the basic concepts of concordances and collocational analysis. It is shown that procedures which work well with smaller corpora and less sophisticated hardware are unsuitable for large corpora. He suggests that indexing, integerization and collocate banks are promising ways of optimising software techniques.

The papers to be discussed now show how corpus linguistic methods can be used to advance linguistic descriptions. Bengt Altenberg's contribution is an illustrative example of corpus linguistic research, it is based on the London-Lund Corpus of Spoken English. Starting from a formal definition of verb-complement constructions (which had to be restricted for practical reasons) and structural statistics, he moves on to describe the interactive functions of these constructions, e.g. *that's right* (agreement), *I'm sure that* (modal), *it's difficult to* (evaluative), *that's the trouble* (retrospective evaluation). There are also lists of fixed expressions like *get in touch with* or collocations, e.g. *have a look at*. The article demonstrates clearly important features of spoken English, namely its interactive character, personal involvement and repetitiveness. Pam Peters has checked statements from usage manuals concerning *like* and *the way* as conjunctions and the distribution of *whom* and *that* vs. *which/who* against the Brown and LOB corpora. She is able to identify regional (AmE/BrE), stylistic (genres) and collocational variables which modify the observations of usage books. Henk Barkema describes a project which studies idiomaticity in English NPs. He introduces useful terminological distinctions, gives a state of the art and reports on various attempts to describe the flexibility of 'received' (i.e. institutionalised) expressions. His investigations show that even a 20 million word corpus such as COBUILD is too small to study flexibility at a large scale. Antoinette Renouf reports on a filter function of the AVIATOR Project; cf. Blackwell's paper mentioned above. In electronic text it identifies words which do not appear in a master word list. Renouf presents the commonest new words in a given period in *The Times* and classifies types of word formation. (The distinction between compounding and combining forms has remained obscure to me.) It is clear that the filter is an excellent tool to describe word formation from the point of view of parole to supplement the langue-oriented approaches predominating in this area. Finally, Willem Meijs reports on research on a special corpus, a database system taken from machine-readable dictionaries: the *Longman Dictionary of Contemporary English* and two Van Dale dictionaries (Dutch monolingual and Dutch-English bilingual). The research interest is to analyse nominal compounds with the help of a computerised knowledge system. The project is an exercise in artificial intelligence using linguistic models to analyse and predict the meanings of nominal compounds.

In all, the volume gives an impressive overview of the new branch of corpus linguistics and informs the reader about the main research centres and directions. But it cannot be overlooked that the discipline is still preoccupied with the establishment of methodological prerequisites such as the compilation of corpora and the development of analytical tools. With a comparatively young discipline as it is, there is naturally a lack of theoretical reflection especially in view of the rapid developments in hard- and software. Corpus linguistics is certainly a kind of (inter-disciplinary) applied linguistics. It may help to solve practical tasks in areas such as the compilation of better dictionaries or grammars and machine translation. But it may also help to satisfy purely linguistic curiosity. It is to be hoped that this aim will find more followers. The prospects are promising because more methodological tools are provided for non-specialist linguists whose research interests are not yet formulated and who should be informed about the new possibilities. This is something the volume can achieve.

Michael Hoey (ed) *Data, description, discourse. Papers on the English language in honour of John McH Sinclair*. London: HarperCollins Publishers. 1993. xv + 175 pp. ISBN 0 00 370947 7.
Mona Baker, **Gill Francis** and **Elena Tognini-Bonelli** (eds) *Text and technology. In honour of John Sinclair*. Philadelphia/Amsterdam: John Benjamins. 1993. xii + 361 pp. ISBN 90 272 2138 3 (Eur)/1-55619-494-3 (US) (alk. paper).
Reviewed by **Pieter de Haan**, University of Nijmegen.

The two volumes under review are both Festschrifts for John Sinclair. From the references in an article in the latter volume, by Gill Francis, I gather that there is even a third, entitled *English in education: Multicultural perspectives.* The editor of this volume, however, is not mentioned. I will start with *Data, description, discourse.*

In his introduction, the editor states that the contributions ‘in this volume reflect ... the range and diversity of interests of the man they celebrate'. This means that almost all the fields of linguistics in which John Sinclair has been active are somehow covered. Most of the papers have some kind of application of concordances as their topic.

In his paper 'Quantitative studies and probabilities in grammar' Halliday describes how he had in the past had occasion to classify grammatical binary systems into those whose two terms occurred roughly equally frequently, and those where one term occurred roughly nine times as often as the other. These findings were based on small amounts of data. He has recently looked at 18 million words of Birmingham data, and found that the tense system (past vs. non-past) is one of the former type, whereas the polarity system (positive vs. negative) is one of the latter type. The second part of this paper describes how the various patterns were recognised in the corpus. Halliday uses a very crude way of correcting figures for spurious occurrences, taking 200 observations at random, counting the number of spurious cases in 200 observations, and then extrapolating from this figure for the entire sample. He estimates that the spurious occurrences do not make up more than 1 per cent of the total number, so that 'the figures are [not] so far out as to invalidate the general picture they present'.

Gross, in his paper on local grammars, attempts to represent sets of similar forms, such as are often found in collocations, metaphors, idioms, etc. and which cannot be related by formal rules such as PS-rules or transformational rules, by the formalism of finite automata. He provides three examples of the working of these finite automata, one of an idiomatic expression, and two of adverbial expressions referring to dates. They fail to impress me, but this is probably because I am not sufficiently into this matter to see where all this should be leading.

Johansson looks into the semantics of adverb-adjective combinations. He distinguishes ten classes (ranging from **manner** and **emphasis** to **value judgment** and **quality and state**.) On the basis of a study of the tagged LOB corpus he shows that many examples do not fit neatly in one of these categories. These may be indications of 'sense change'. He proposes more detailed collocational studies on the basis of larger corpora, such as the Bank of English, to gain more insight into the development of sense changes and, at the same time, the complexity of meaning and collocational patterns in adverbial modification in general.

In his article on aspectual verbs in dictionaries, Gerhard Leitner discusses the treatment of the verbs *begin* and *start* in three learner's dictionaries: the *Oxford Advanced Learner's Dictionary* (OALD 1989), the *Longman Dictionary of Contemporary English* (LDOCE 1987) and the *Collins COBUILD English Language Dictionary* (CCELD 1987). He points out the need for a system, following Sinclair, in which grammatical and usage information, in so far as it is relevant for certain senses, is incorporated into the dictionary definitions, admitting at the same time that this is something that will be difficult to achieve.

Stubbs and Gerbig, in their article, use a lexical analysis of a geography textbook to find an answer to the question in which way the world is talked about. They look at the linguistics of representation, lexical density, change, causation and agency (the relationship between transitivity and ergativity). This leads to a discussion of comprehensive text analysis, which is a discussion of the philosophy and the methodology of this type of analysis, with a plea for an integrated software environment enabling researchers to access various information and comparison routines, such as concordances, word frequency counts, etc. The textbook, finally, is established as a genre in its own right, as textbooks are 'institutionally sanctioned versions of knowledge: what is believed to be worth passing on to pupils'. This genre is not represented in either Brown or LOB. One conclusion that is drawn on the basis of the study of verbs in this particular textbook is that in it the world is represented as one where human beings are largely absent as responsible agents, where processes take place spontaneously or are caused by other abstract processes.

Malcolm Coulthard exemplifies some methodologies used in forensic linguistics, based on a couple of real life cases, and the role of corpus linguistic research in this. He says that linguistic evidence in court will usually be probabilistic and that it is highly unlikely that any method of 'linguistic fingerprinting' will be developed in the near future, despite claims of Morton (reference to whom is missing). By comparing a general corpus with a specialised corpus he is able to show that the postpositioning of 'then' (as in *He then...*, rather than *Then he...*) is a very reliable distinguisher of police register. The occurrence of postpositioned 'then' in a statement could be taken to be indicative of some intrusion of policemen's register in the statement.

Angele Tadros presents us with an investigation of the interactions in written texts through text averral (i.e. making no explicit reference to another source) and text attribution (where explicit reference to another source **is** made). She provides examples of both, taken from three different texts, showing that the purpose of the text (the reader-orientation) plays an important role in the choice of averral or attribution. Thus, in an introductory textbook the author may want to be seen as the authority in the field, which means that there is hardly any attribution. In a more advanced textbook there is more attribution, but of a ‘mitigating’ kind, its purpose being to show the readers that previous studies on the topic are not to be discarded outright. In a book which has the status of a research report, finally, attribution abounds, if only for the author to show the readers how previous work has certain inadequacies with respect to the data to be described. The pedagogical implication of this study should be clear. Students should be told to be aware of the ‘various voices they hear within the same text’. Likewise, they should learn to signal clearly when they have switched from expressing their own views to reporting and vice versa, if only to avoid being accused of ambiguity, or even plagiarism.

Hoey makes the case for the **exchange complex** in the rank scale of classroom interaction. He argues that the concept of **transaction** is all right as an indication of organisational aspects of (classroom) discourse, but not with respect to its structural status. For the latter the exchange complex is a more adequate concept. He draws a parallel between the structure of classroom discourse and text structure:

sentence/clause complex	exchange complex
clause	exchange
group	move
word	act

Various patterns of exchange complexes are discussed and exemplified: branching exchanges, converging exchanges, subordinate exchanges and coordinated exchanges, and combinations of these. It would seem that this structure enables the exploration of interactive development of a discourse.

Carter makes the point of the desirability of teaching language learners knowledge **about** language, rather than just knowledge **of** the language. The latter is advocated by proponents of communicative language learning (language acquisition), whose aim is to teach an intuitive knowledge (knowing **how** to use a language fluently). They oppose language learning which is essentially focused on the structures of a language in an explicit way, as it means that a considerable amount of meta-language is needed and this makes life harder on the learner. Carter proposes to integrate knowledge about language and knowledge of language. His argument is that learners may be quite satisfied with only knowledge of the language for communicative purposes, but will need knowledge about the language in certain cases, for example to appreciate advertisements, puns, or poetry. He distinguishes three parameters of language awareness: one of **form**, one of **function** and one of **socio-cultural meaning**, though the difference between the second and the third is not clear to me.

Kapland, finally, is the odd one out here. He makes an inventory of the language situation in New Zealand. After giving a survey of the various languages in New Zealand, and discussing government policy with respect to language planning, he concludes that the Maori language is marked by the three conditions that may signal language death. They are 1) parents do not engage in intergenerational transmission of the language - in many cases only grandparents are capable of passing on the language; 2) the language loses registers to another language, in this case English; 3) younger individuals are drawn away from the community by economic pressures.

Text and technology consists of three sections: 1) Spoken and Written Discourse, 2) Corpus Studies: Theory and Practice, and 3) Text and Technology: Computational Tools. Within the context of this review I will not discuss the first section. Section two opens with an article by Gill Francis, which she calls ‘the corpus driven approach to grammar’, but which would have been more appropriately called ‘the Cobuild approach’. She discusses the characteristics of the Cobuild approach to a new descriptive grammar of English. There is first of all, the fact that they provide only real examples. Secondly, the grammar is aimed to be specifically non-contrastive, by which she means that the grammar will show how ‘each item or structure is used in its own right, rather than as compared with members of the same or a contrasting paradigm’. No argument is given for this choice, but ‘the development of this standpoint is high on the agenda’. Does this mean that they are still working out why they want to be different from Quirk *et al.* in this respect? I do not know. I could imagine, though, that anyone consulting a descriptive grammar of the language would benefit as much, if not more, from reading precisely about such a comparison. The third point is that the grammar will be completely data-driven and not access any linguistic intuition at all, as this ‘is notoriously unreliable’. The corpus is the only authority. The fourth characteristic is that it is essentially a lexical approach. Francis points out, taking the example of anticipatory *it* used as direct object, that the two verbs that are by far the most frequently complemented by this construction are *find* and *make*. Likewise, there are a limited number of adjectives that can follow *it* in these constructions. As such this is different from what Quirk *et al.* have to say about it. They mainly comment on the form of the clause that *it* is used as a stand-in for. Personally I feel that Quirk *et al.*'s approach looks more like grammar to me, as it accounts for the structure of the language. The same goes for the other example Francis mentions, that of appositive *that*-clauses. She breaks down the category of head nouns into six groups, each with their own specific meanings and combinations. No mention is made of other forms of appositive clauses, such as infinitive clauses, nor is a parallel drawn between appositive *that*-clauses and cases like *the question whether...* An extreme consequence of this approach is when the collocation *for the simple reason that...* is said to be 'so predictable that the phrase can be seen as a single-choice chunk..., in fact it no longer makes sense to think of *reason* as the head-noun within a nominal group' (p. 153). No alternative suggestion is offered, though. To me this looks like a confusion of syntactic with semantic description. A similar point was made by Collins (1993) in his review of the *Collins COBUILD English Grammar*. It is a pity that Francis has missed a chance of an interesting discussion. On p. 139 she writes:

> A third defining characteristic of the grammar we are compiling is that it will a be data-driven grammar of English. It might be supposed that this is a principle held by corpus linguists in general, but that this is not the case is evidenced by arguments like that of Aarts (1991:45), who suggests that the corpus linguist make up his [sic] grammar on the basis of his 'intuitive knowledge of the language and whatever is helpful in the literature',...

It should be noted that the [*sic*] in this passage is not mine, but occurs in the text. I have been brought up to believe that you use [*sic*] in a quotation to tell your readers that this was what you really found in the text, unlikely though it might seem, usually signalling a typing error or an error of fact in the original text. Francis uses [*sic*] in her own paraphrase of a quotation. She apparently wishes to stress the point that grammars should be dictated by the corpus. In taking this position she apparently skips 30 years of linguistic history. Transformational grammar has shown the value of intuitive data, if only by the introduction of the notion **ungrammatical**. Aarts writes:

> The (corpus) linguist first writes a formal grammar for some well-delimited part of the corpus language, for example, the noun phrase. This grammar is written on the basis of the linguist's intuitive knowledge of the language and whatever is helpful in the literature.
> (Aarts, 1991: 45)

Francis, on the other hand, says that intuitive data are wrong and should not be used. Incidentally, she fails to mention that this was a passage that Aarts quoted from an earlier paper (Aarts, 1988). Moreover, the reference to Aarts is incomplete. The correct reference can be found below.

Bill Louw shows how a study of collocations, revealing **semantic prosodies**, i.e. the way in which habitual collocates of a certain word or phrase can colour it, can help to determine irony in a text.

Partington studies intensifiers form a perspective of language change. He looks at a limited number of adverbials, and compares their use in the past with their use in the present, showing that most of them have a more limited syntactic range today than in the past, which has meant a change from modal to intensifying use.

Tognini-Bonelli discusses the specific discourse functions of *actual* and *actually*. The adjective *actual* is said to qualify as a device of postural change on two grounds. 1) It performs an act of self-reference, i.e. refers to a preceding unit of discourse, and in a way redefines it, usually from a broader, consensus-based view to a more specific, unexpected and implicitly privileged interpretation. 2) It offers a different interpretative angle on the subject-matter. *Actually* also often indicates, unexpectedness; given an implicit or explicit 'norm', it will indicate and highlight a deviation from, or contradiction of, this norm. It is also a means of self-correction and mitigation (in order to avoid embarrassment) or challenge (to make a contradiction run more smoothly).

A really nice application of concordances is presented by Kirsten Malmkjaer, who compares a number of translations of Andersen's fairy tales, in order to prove, be it in a very roundabout way, that translators might produce higher degrees of equivalence if they are translating out of their mother tongues than if they are translating into their mother tongues. The reason for this is said to be that near-native L2 speakers are almost like natives in formal features (production - intuitive judgement about part of speech), but not in functional or cognitive aspects of grammar (interpretation - intuition about use). It is suggested that somebody who is a near-native, but frequent and competent user of L2 will be sufficiently aware of the uses in his/her own native language, but competent enough in L2 to bring about more adequate translations. A point that is mentioned in passing, viz. the discrepancy between speakers' intuitions about their language and frequency data gathered from corpora, is not satisfactorily dealt with.

Baker mentions the low status accorded to translated texts and pleads for translation studies as an academic discipline in its own right. She mentions a few translation universals that corpus studies could help to find. They include such things as explicitation; disambiguation/simplification; conventional grammaticality; avoidance of repetition; exaggeration of features of the target language; specific distribution of features in translations that are neither typical of the source language nor of the target language. She ends with a rather lame statement that 'a suitable methodology and a set of very powerful and adaptable tools are now available from corpus linguistics' to achieve these goals, without even hinting at what these tools are. One suspects that they are the good old concordancing programs.

The papers in section three are more interesting, as they are about methodology. Coniam discusses some results of a partial parser, one of a suite of tools described by Sinclair, aimed at determining the type of boundary marker a space between two words is; a word, group or clause boundary marker. This is done on the basis of a first-order Markovian model, which takes only the word itself and the words on either side to make its decisions. Parsing takes place without reference to explicit grammatical rules. The idea is that the outcome of this process would be passed on to the next tool in the line, which would, again, perform a small, finite task, and pass on the result to a third, and so on. Only the crudest groups and clauses can be distinguished in this way. It is suggested that more complex groups and clauses cannot be distinguished without recourse to explicit grammatical rules. This leaves me with the question what the advantage is of having partial parsers, which can apparently only perform adequately with reference to explicit grammatical rules, over a system which attempts a full parse right away, seeing that the explicit grammatical rules are necessary anyway.

Clear discusses the phenomenon of stereotyping, i.e. the process of words slipping out of their appointed range to form particular attachments. Ideally in a corpus the search for interesting word pairs should have a high precision (not yielding uninteresting combinations) and high recall (giving indeed the combinations that we are interested in). Usually there is a trade-off between the two which is no more than a weak compromise. He discusses methodological aspects of the current version of the collocate program. They are first of all the span, which is by default taken to be a five-word window (two words left and right of the keyword). Next, the frequency of co-occurrence, where a threshold of three is used, which means that word pairs which occurred fewer than three times are discarded. The drawback of this is that new collocates are not easily identified. Third, there is no default lemmatization of keywords, and none at all of the collocates. However, this is not considered to be a serious problem as it has been found that one of the inflected forms will appear as a collocate. Finally, there are two measures of significance, the MI-score and the T-score. MI indicates the strength of association between two words, whereas the T-score indicates the confidence with which association can be claimed to exist. The difference between the two measures is convincingly exemplified.

Nakamura has studied the use of public verbs, private verbs and suasive verbs in four large sub-corpora of the Bank of English. He uses Hayashi's Quantification Method Type III to rearrange the rows and columns of a data matrix containing the four subcorpora and the frequencies of 149 verbs. A look at the reference list tells me that he has used this method a number of times to do similar things with respect to the genres in the Brown and LOB corpora. The results of this procedure can be plotted in a series of two-dimensional figures, showing how the various sub-corpora tend to coincide with particular verbs. This provides an objective means of testing our intuitive knowledge of the distribution of texts and corpora. I wonder whether the method might also be used to identify text types within each of the four subcorpora, which could be used to assess the homogeneity or heterogeneity of the sub-corpora. Nakamura does not go into this.

Barnbrook is developing a parser for the functional components of the Cobuild definitions, the outcome of which could be used for a wide range of natural language processing applications. A very convenient element in the definitions is the use of bold type to highlight the headwords. The codes turning bold type on and off usually divide the definitions in clearly recognisable chunks that have their own patterns. It appears that in the definitions a limited number of patterns is used. These are discussed and exemplified. The ultimate goal is the creation of a dictionary database, which could be used to generate an automatic thesaurus, to refine the dictionary explanations and even to turn monolingual dictionaries into some kind of bilingual dictionaries.

Allan, finally, discusses the use of hypermedia networks for the study of intonation in ESL/EFL. She takes as a basis a Vygotskyan theory of learning, which says that learning to solve problems takes place through interaction with others, which later leads to the adoption of similar strategies in solving problems alone. It can be argued that the teacher's role in this process is to diagnose learners' capabilities in any area, and to be prepared to adopt different strategies in providing support to different learners and to have resources which facilitate shifting goals. This is where the hypermedia network comes in, with its random access to any node in the network. The application of such a network is exemplified in a prototype version of a program designed to help students to master discourse intonation.

Of the two volumes Hoey's *Data, description, discourse* strikes me as more clearly honouring John Sinclair. Each of the contributors has done this in his or her own way, with the result that there is far less coherence between the papers than in the other volume, which has a more obviously thematic set-up. But perhaps coherence is something that cannot or must not necessarily be expected from festschrifts. In *Text and technology* I cannot help feeling sometimes that the editors may not always have been fully aware of the conflicting implications of the various papers. Thus Francis advocates the strictly data-driven approach, making the corpus set the norm, whereas Louw and Clear, to mention but two, show that language users will go off beaten tracks, as they are inventive and creative, which is something that Aarts was fully aware of when he formulated the passage in Aarts (1988). Different readers will undoubtedly react differently to the papers in the two volumes. I have tried to indicate my preferences.

References

Aarts, J. (1988). Corpus linguistics. An appraisal. Paper read at the Fifteenth International Conference on Literary and Linguistic Computing. Jerusalem. June 1988.

Aarts, J. (1991). Intuition-based and observation-based grammars, in Aijmer, K. & B. Altenberg (eds.): *English corpus linguistics*. London & New York: Longman.

Collins, P. (1993). Review of Sinclair, J., G. Fox *et al. Collins COBUILD English Grammar*. In: *IRAL* XXXI/2: 161-167.

Jane A. Edwards and **Martin D. Lampert**, (eds.). *Talking data: Transcription and coding in discourse research.* New Jersey: Lawrence Erlbaum, 1993. pp vi + 325. Reviewed by **Gerry Knowles**, Lancaster University, U.K.

Linguists have always asserted the principle of the primacy of speech over writing, and they have nearly always in practice studied speech through the written language. Even in corpus linguistics, studies of spoken data have in many cases concentrated on that arbitrary subset of speech that can be written down in standard orthography. It is therefore most encouraging to find a book which tackles the problem of representing the wide range of phenomena in speech that cannot be written down. The contributors to this book are well chosen not only in that they come from different disciplines, but they also have extensive practical experience of the problems under discussion.

The book is divided into three parts, dealing respectively with transcription, coding and resources. Part III is a useful reference work, in which Jane Edwards reviews available corpus resources, including treebanks, phonetic databases and corpora of languages other than English. This review will concentrate on parts I and II.

The agenda for the book is set in the opening editorial chapter by Jane Edwards entitled ‘Principles and constrasting systems of transcription’. This is in the first place a valuable survey of what scholars have done to date. However, the very clarity with which she presents current practice and the assumptions that lie behind it opens up some fundamental questions. First, what is the relationship between the transcription and the original events? The categories marked in the transcriptions shown sometimes represent a subset of the data (e.g. when pauses are measured in seconds) and sometimes an interpretation of the data (as when a syllable is said to be ‘stressed’). Are these categories formal or functional, or a bit of both? Secondly, how does the transcriber decide on the set of categories to be used? The categories are assumed to be Aristotelian rather than prototype categories (p5). However, it is only in invented classroom data that categories are so well defined that problems of classification never arise, and one of the familiar problems of transcribing natural data involves the fuzzy boundaries between categories. How is fuzziness to be handled? Thirdly, there is an emphasis on the presentation of transcriptions on the printed page. More fundamental, to my mind, is the problem of defining what is to be annotated, and how it is to be stored in computer readable form. Presentation on the page – and, for that matter, how the transcription is input – is a separate issue. It is worthy of note that the TEI (Text Encoding Initiative) is mentioned (p10) but in passing. The same is true of the TOBI (TOnes and Break Indices) transcription system (p14) which surely merits a more detailed discussion. The tone is set for a book which deals with what scholars actually do at the present stage of understanding, not what they might do in a theoretically perfect world.

The first requirement of a transcription system is to represent in a consistent manner the data on which analyses are to be based. In chapter 5, 'HIAT: a transcription system for discourse data', Konrad Ehlich describes the Halbinterpretative Arbeitstranskriptionen system. The basic system handles words, prosody and turntaking; there is also an elaborate system to add prosodic detail, and to deal with non-verbal communication and actions. An attractive feature of HIAT is that it is designed for use in a computer environment. Not only does it provide facilities for inputting data at the keyboard and for the automatic formatting of the input, but the alignment of different kinds of annotation make possible the retrieval of the necessary information for discourse research. The theoretical assumptions behind the categories annotated, as far as they are made explicit, are likely to receive the assent of linguists and phoneticians.

In chapter 3, 'Outline of discourse transcription', Jack Du Bois, Stephen Schuetze-Coburn, Susanna Cumming and Danae Paolino make a comprehensive survey of the kind of information that is worthy of annotation in discourse. Sufficient material for a book is here compressed into 45 (actually very readable) pages. Annotations are of several different kinds, some phonetic, some phonological, and some representing discourse structure. Some are mixed, e.g. the symbol for latching (p63) simultaneously measures the phonetic duration of the pause as zero, and marks the nature of the transition between speakers. In some cases, the theoretical status of a category is left rather vague. This applies to the key categories intonation unit and accent unit, which are essentially units of discourse but are defined in apparently phonetic terms. The set of attributes used to identify the intonation unit (Du Bois, personal communication) makes it identical to the tone group of the Lancaster/IBM Spoken English Corpus (p285) which is a very much smaller unit of discourse. This is a serious shortcoming, which has consequences in chapters 2 and 9.

Intonation units also figure in Wallace Chafe's chapter 2 'Prosodic and functional units of language'. Chafe interprets these units in cognitive terms and relates prosody to the flow of information in discourse. Intuitively I feel sure that Chafe is on the right track, but the descriptions are not sufficiently explicit for the claims to be tested. This is frustrating, because the clarification of vague concepts like the intonation unit (and its British cousin, the tone group) is the precisely the kind of problem that can only be solved by the analysis of carefully annotated speech corpora and databases. The overlap with chapter 3 could have been reduced and the discussion of the annotation of information structure expanded. Chapter 2 would also be more logically ordered after chapter 3.

In chapter 4, 'Transcribing conversational exchanges', John Gumperz and Norine Berenz concentrate on the interpretative evaluation of speech events in interactive situations. They start with the problem of how to characterise the relevant events at the appropriate degree of abstraction for the purposes of conversation analysis. However, the events that they transcribe are not properly defined and do not relate to the problems they set out to solve. These events are described as though they were phonetic events, but reading the transcriptions requires more the skills of the actor than those of the phonetician. *Voice qualities* (p108) include hi, lo, ac, dc, f, ff, p, and pp. These are not voice qualities but matters of pitch range, tempo, and loudness. *Fluctuating intonation* is a category which includes the fall-rise and the rise-fall tones, but there is no criterion by which these can be regarded as a natural class. Popular spelling is used to represent pronunciation, e.g. *ahma git* (p97); phonemic transcription is referred to on the same page but for some unspecified purpose. This is a disappointing chapter, because it deals with an important set of problems in an amateurish way.

Most coding systems described as 'transcriptions' retain the horizontal format of the orthographic text. Lois Bloom's chapter 'Transcription and coding for child language research' reports work based on video recordings, and starts instead with a view of the data as a sequence of frames. The data is presented as a flatfile (p159) with an explanatory header, and using the frame numbers as record identifiers. This approach separates the problems of storing annotated data and presenting it to the human reader. The result may look odd to those unaccustomed to seeing data presented in this manner, but perhaps it is time for linguists to wean themselves off book format, and to consider the organisation of data as an interesting problem in its own right.

Part II opens with a key chapter, 'Structured coding for the study of language', in which Martin Lampert and Susan Ervin-Tripp follow a trail through the coding process. They first examine the construction of a coding system, including what to code and the relationship of codes to theory. In order to implement the system, the coders have to be trained and provided with criteria for the use of codes, and as an illustration a very explicit flowchart for the encoding of control acts is given on p191. The system has to be evaluated, to test the reliability and consistency of coding. They point out that measures of agreement between transcribers should exclude cases of agreement due to chance. (This may be perfectly obvious, but I shall be quietly changing the procedure I myself use to measure agreement!) The chapter ends with a discussion of ways in which codes can be used with packages to retrieve and analyse the data.

An important problem of design is identified by Dan Slobin in the next chapter (‘Coding child language data for crosslinguistic analysis’), namely whether codes should apply to linguistic form or function. The problem is illustrated with relative clauses in English and Turkish. The Turkish equivalent of ‘the potato that Hasan gave to Sinan’ can be glossed as ‘the potato of Hasan’s giving to Sinan’; in other words although Turkish does not have a formal device that can be labelled *relative clause* it certainly has a functional equivalent. As Slobin points out, functional equivalents are also used in informal English. The comparison of languages – or indeed registers – requires coding at the functional level as well as the formal level. This is a fundamental point which applies also in other areas, including prosody.

In chapter 9, ‘Representing hierarchy: constituent structure for discourse databases’, Du Bois and Schuetze-Coburn tackle a fundamental problem for linguistic databases. Most databases are designed on the assumption that every record is independent of every other record. This is fine for CD collections and alumni address lists, but not for texts, which are ordered in highly complex ways. Not only does a word or phrase fit into a syntactic hierarchy, but it also fits into other structures (including prosody and perhaps information) which are not even properly understood. The representation of these structures, and the formal linking of different kinds of structure, is an extremely interesting and currently relevant research problem.

Du Bois and Schuetze-Coburn present a conventional syntactic tree (p226) and reproduce it (p227) in horizontal format as a labelled bracketing added to the orthographic text. They find the result unreadable – as it always is – and at this point they retreat from the problem, regarding it as insoluble. But this is to confuse the organisation of the data with its presentation. The problem of organisation is certainly not insoluble. Every item in the tree is a member of the item in the next level up. At the bottom level this is a paradigmatic relation, e.g. weird is a member of the class of adjectives, and this is handled by making the grammatical tag a field of a word-sized record. Other relations are syntagmatic, e.g. the adjective in he is weird is a member of the VP, and the VP is in turn a member of the Sentence. These consistent relationships (of X as a member of Y) can be stored in a separate table. Admittedly this entails abandoning the book format, but with data of this level of complexity, it is naive to try to cling to it in the first place.

In the sample database file, the text is divided into records of the size of a *group* which earlier in the text is ambiguously described as a unit of syntax or prosody, but here used as a prosodic entity, as a constituent of the intonation unit. This structure is built into the record identifier, e.g. HYPO.73.3 is the third group of unit 73. (It is also tagged as a verb, which suggests it is a syntactic entity.) Before we go any further, we must note that this database can only be used by someone who accepts that the transcription on which the file is based represents God's Truth on the prosody: if any change is made in the transcription the file has to be re-compiled.

Syntax is represented in a rudimentary form, e.g. by using # to mark the end of a clause. It is clear, even in this small sample, that # occurs at the end of intonation units. But since the syntactic boundary will almost certainly have been used by the transcribers as one of the cues to recognise the end of the intonation unit in the first place, this is hardly surprising. Since the ends of other groups are not annotated at all, it is impossible to make any non-circular inferences concerning the relationship of prosody to syntax. This database would have been much more interesting if the record had been based on the orthographic word – which everybody can agree on most of the time – and the word related independently to syntactic structure on the one hand and to the prosodic transcription on the other. It would then be possible to link in an independent annotation of the information structure, and test some interesting hypotheses concerning the role of prosody in discourse.

Taking the book as a whole, the questions being raised are timely and essential for the analysis of large corpora and databases of spoken material. The answers offered here – despite some false starts and blind alleys – point to interesting developments in the future. These will require a more rigorous approach to the data, both in the definition of annotations, and in linking related pieces of data. In particular, the design of databases must be determined by the nature of the data, not by the need to present it on the printed page.

Finally, who is this book suitable for? I hesitate to agree with the editors that it is suitable for undergraduates. On the other hand, for researchers and research students working under supervision, who have to decide how to deal with bewilderingly complex speech data, it will be an excellent training manual.