REVIEWS

Jan Svartvik (ed.). *Directions in corpus linguistics: Proceedings of Nobel Symposium 82 Stockholm, 4–8 August 1991.* Trends in Linguistics Studies and Monographs 65. Ed. Werner Winter. Berlin and New York: Mouton de Gruyter, 1992. xii, 487. ISBN 3-11-012826-8. DM 218,00. Reviewed by **Ian Lancashire**, University of Toronto.

The thirty-four "Ladies and gentlemen, dear colleagues and friends", whom Sture Allén, Jan Svartvik, and their colleagues gathered together as guests of IBM at its Nordic Education Center on the island of Lidingö in August 1991 included the founders of modern corpus linguistics but also some of the ablest young minds in the subject today. Five Britons, five Scandinavians, six Americans, two Australians, and a New Zealander gave 19 papers, over five days, on the history, theory, design, development, exploration, and application of diachronic and modern synchronic English and Swedish language corpora. Directions in Corpus Linguistics differs from annual ICAME volumes, which give researchers an opportunity to publish the fruits of current projects, and from the monographs that grow from them. Jan Svartvik has goals broader than these. Directions aims at placing corpus linguistics, as a subject, at the heart of scientific research in language studies no matter where that study occurs – whether in departments of cognitive studies, computational linguistics, languages, linguistics, and phonetics, or in industrial laboratories like AT&T or IBM – and at charting, through consensus, research strategies towards that goal. In this Svartvik has done well. This fine collection of papers deserves a wide readership both in the language industries and among those on whose shoulders rests the important task of defining the research objectives, methods, and applications of corpus linguistics for the benefit of society at large. The contributors to this intellectually rich book earn its importance. What makes their work seminal is its perspective: the identification of directions.

Born in the mutual friendship and respect of colleagues well-established internationally in their various disciplines and perhaps with less to prove than junior researchers, invited conferees will sometimes skimp on presenting new knowledge in their fields, but this is not true of *Directions*. Its papers often present new historical, theoretical, or experimental material and expose it to fresh thinking. Also, thanks to Jan Svartvik's

abilities as a host, the Nobel symposium worked its speakers hard. Many had a commentator examine their ideas. General niceties notwithstanding, the critics gave little quarter, and the action on the court proved brisk, as some excerpts from the exchanges show: "a polemical overstatement", "should be applied with some caution", "and "Is this always the case?" No one was skewered, for consensus existed on many issues, but the general comfort level seems to have dropped at times.

Two innocent-enough-looking poems opened and closed the proceedings, ones by Lars Huldén ("You know, my dear Teophilus, that / in Heaven there is a concordance") and the cryptic pseudonym Anna Kerr-Luther ("~~The Purposes of Corpuses~~ / ~~The Purpora of Corpora~~ / How to Dispose of the Body / Or / The Corpus-Maker's (Di-)Lemma"). Sture Allén translated the first poem to remind his listeners that their words, "as well as / context enough for assessing / their inward sense", will be permanently on file for the study of both forgiving angels, those with everyone's best interests at heart, and of devils, who have something else at stake. (As an ICAMEr once said to me as I was about to give two consecutive visiting talks to an unknown interdisciplinary audience, "Now we have a little pressure". There is a delicate balance between angelic fondness for one's colleagues, and alertness to the traces of ancient demons in their humanity). It is at the end of the collection that the editor prints Kerr-Luther's delectable poem, found "on a table among the debris" at the end of the conference. Through its puns ("Svart(vik) Boxes") and malapropisms ("a data-base corpuscular"), this ditty advises corpus-builders of the 1990s in a way that earns it the praise of the editor as capturing, "succinctly, accurately and elegantly – the spirit of the whole Symposium":

> Don't rely on introspection, on the magic and the mystical,
> Don't hope to find a software pack of strategies heuristical,
> Just build yourself a corpus (or some corpora) statistical!

These innocuous sixteeners define *statistical* analysis as the key method in the future science of corpus linguistics. Does the last word say it all? Is the symposium a cautionary tale that the "Hidden Markoff Process / [that] Lurks beneath your text" is the direction for corpus linguistics?

Jan Svartvik's opening paper, "Corpus linguistics comes of age", discusses the *status quo* and raises doubts that any one direction lies ahead. Defining corpus linguistics as "the use of large collections of text available in machine-readable form" (7), Svartvik surveys reasons why researchers opt to use a corpus for the purpose of generalizing

about language behaviour. Unlike introspection or elicitation, linguistic corpora shared among researchers make possible for them in public (and without having to be a native speaker) to verify all results, to turn to the same data source repeatedly for many kinds of language features, to compare studies of different features, and to analyze language across time and across registers, tasks not well served by other methods. Corpora also attract researchers from many fields, especially outside empirical linguistics. A theorist can test "rule systems that have built-in predictability"; a language teacher can derive examples; a literary historian can study style; and a software developer can base a grammar-checker on corpus data. Svartvik predicts future corpora of massive size (100 million words and more) and different structure (the monitor corpus, in particular), both widely and inexpensively available on international networks for teaching and research. One word that never occurs in his chapter, however, is *statistics* or *statistical*. Svartvik remains sceptical whether any machine can do better than "the human mind", that is, "soft human intuition", in understanding corpus data. He especially warns against the loss of hands-on familiarity with a corpus that comes with using heavily encoded versions transcribed in ways that make language "a kind of canon and context-free object". He has impromptu speech in mind here.

W. Nelson Francis's "Historical conspectus B.C." ('before computer', but 'Brown Corpus' is an amusing pun) discusses how the *status quo* evolved, at least to the date of the publication of the Brown Corpus. Dividing corpora into lexicographical, dialectological, and grammatical varieties, Francis shows that the making and analysis of representative linguistic corpora thrived long before computers were available. He describes, among other works, the 150,000-sentence corpus behind Samuel Johnson's dictionary (1755), the eleven million paper slips of the *Oxford English Dictionary*, the 835-page corpus of English dialect material for the work of Alexander J. Ellis on *The Existing Phonology of English Dialects* (1889), the over 400,000 items in Harold Orton's Survey of English Dialects (1962–71), and Randolph Quirk's *Survey of English Usage*. Although these corpora were made by pioneers, Francis notes that their works are being computerized today, joining computerized corpora of the past twenty years. Automating these "painfully" hand-made monuments can be managed only because they were erected systematically in the first place. Francis shows that, even in corpus linguistics, there is little new under the sun, although readers of this book are still waiting for the first mention of statistics.

It first appears, of all places, in the first of four theoretical papers, Charles J. Fillmore's "'Corpus linguistics' or 'Computer-aided armchair linguistics.'" Unrepentantly "an armchair linguist who refuses to give up his old ways", thinking about sentences, Fillmore nonetheless says, "... every corpus that I've had a chance to examine, however small, has taught me facts that I couldn't imagine finding out about in any other way" (35). He then illustrates his point by discussing his research on two words, *risk* and determiner-less *home*, the former done with Beryl T. Atkins on 1743 sentences collected from a 25-million-word corpus from the American Publishing House for the Blind, and the latter on 450 sentences from the 8-9-million-word *Wall Street Journal* section of the DCI corpus. The first example suffices to make Fillmore's case. By dwelling on his corpus citations, he and Atkins learn that we use *run a risk* in circumstances "where there is the possibility that some harm will occur, but not necessarily as the result of someone's action", but we *take a risk* on recognizing that something we do, consciously or not, puts us in danger. Thus, corpus linguists run (not take) a risk by leaving home without an umbrella but take (not run) a risk in hiring others to proofread their texts. The American Publishing House corpus included citations that would have occurred to neither researcher and that alerted them to a problem with current dictionary definitions, yet it offered no examples in which the substitution was actually impossible. Because this anti-substitution constraint was the main result of their research, armchair linguistics carried it to conclusion where the corpus could not. And hence Fillmore's insight: "there are no corpora of starred examples: a corpus cannot tell us what is not possible" (58). Properly to use corpus output, in his mind, means to "sit down and stare at the examples one at a time to try to work out just what is the intended cognitive experience of the interpreter, what are the interactional intentions of the writer, and so on" (59).

Fillmore approaches his corpora as snapshot albums of utterances, verbal or written, by many individual minds, albeit operating collectively sometimes in ways they do not understand. For others at the symposium, however, corpora represent a faceless (dis-minded?) mass of text, and corpus linguistics a field where speaking of linguistic features in terms of intentionality does not make sense.

In "Language as System and language as instance: The corpus as a theoretical construct", M. A. K. Halliday gives us a powerful, clear, and eloquent rationale for the view that an important new direction in corpus linguistics lies in probabilistic modelling of grammar on the basis of

evidence in very large corpora. Halliday begins this essay with some apt personal academic history, particularly about his "Nigel" grammar, "a network of 81 systems each with a probability attached to the individual terms" that, when implemented by a random language generator that originally output utter garbage, suddenly "produced garbage that now actually looked like English" (65). It is little wonder, then, that Halliday has come to believe that "frequency in the corpus is the *instantiation* (note, not realization) of probability in the grammar" (66). His paper proposes, at its centre, seven questions in the statistical analysis of corpora for the near future. Anyone at work in corpus linguistics is well advised to pay close attention to pp. 67-76, where Halliday explains these questions, for this section clearly explains the relevance of statistical analysis to linguistics in ways that those not in the know will find helpful.

Let me summarize the seven demands to be made of corpora. First, Halliday wants to know whether the relative frequencies of what he calls "low-delicacy" grammar systems follow a general probability pattern. For example, do indicative and imperative moods, or active and passive voices, or singular and plural number occur about equally or vary by a sizable amount, say ten to one, and are there no other probabilities at work but these two? Second, to what extent can registers be discriminated by "variation in the setting of grammatical probabilities"? Halliday believes that any register is "a syndrome of lexicogrammatical probabilities" (68). Third, he wants to discover whether the probability for selecting one term in a grammar depends on which term occurred previously. This interconnectedness of probabilities is what Halliday, and Anna Kerr-Luther in her poem, mean by a Markov process. Fourth, he proposes certain measures now available in corpus analysis that would help us identify where, and by how much, complexity – of nominal strings, ranking clauses, etc. – increases dynamically in texts. Fifth, Halliday extends his third question about conditional probabilities by asking how one grammatical system (not just one term) favours another in the same context (not just subsequently, because systems like mood and voice are chosen prior to sentence formation). The sixth question, less easy to grasp, has to do with historical linguistics. Halliday suggests that at an early stage in history some grammatical system, say the balanced pair of direct speech and indirect thought, might have split up into its present four parts, direct (quotation), indirect (report), speech, and thought. Yet while these might still retain their original associations, additional pairings would be possible, e.g., direct (quotation) and thought.

Halliday argues that gradually mounting probabilistic complexities of this type explain how many new meanings are created over time. The seventh question treats recursion, which Halliday suspects is chosen by someone as an option in speech or writing about ten percent of the time. He asks whether this could be the "single pattern of frequency distribution covering all kinds of ‘marking’".

Halliday’s use of statistics to comprehend language is not deterministic. He argues that, although children learn how to speak and write by building up "a probability profile" of both lexis-grammar (68, 76), they and we always retain the freedom to choose to form an utterance so that it violates probability. He cannot understand why people think that assigning probabilities to linguistic features threatens "the freedom of the individual" (76). Halliday also believes that a statistician like himself, and an "instance-observer" (as he calls the armchair linguist), are studying the same thing; it is just that the former behaves like a climatologist, and the latter like a weatherman.

Svartvik’s introduction ("Corpus linguistics comes of age"), the first three papers by Francis, Fillmore, and Halliday, and Randolph Quirk’s postscript ("On corpus principles and design") are distinguished by having no commentator. Svartvik earns his peace by playing the Host in this modern *Canterbury Tales*. (It is Randolph Quirk who cites John Dryden’s verdict on Chaucer’s works when he reviews the symposium in a postscript essay: "’Tis sufficient to say ... that here is God’s plenty".) For some, Quirk’s genial review of the proceedings may serve the purpose better than mine, intermingled as his comments are with an unveiling of the 100-million-word British National Corpus project (on the Advisory Committee to which, alone of all those at this invited conference, he sits). This undertaking dwarfs every other corpus discussed in the volume, especially the original Brown Corpus project, whose planning meeting he attended in 1963. It is hard to imagine who could comment on Quirk’s postscript, except to say that the BNC is, in itself, a guideline if not a direction for the entire discipline, since it encompasses most of the objectives and methods discussed at the symposium. Summarizing its conclusions, Quirk agrees with Geoffrey Sampson’s remark that we "still have a long way to go" while adding, "What a long way we have come". This deft compliment characterizes Quirk’s manner. Consider the following statement:

> ... my colleagues and I have demonstrated that sophisticated elicitation procedures could establish for one’s own language statistically

> significant generalisations which resisted introspection and could scarcely be imagined as emerging from corpus scrutiny alone (though corpus data could often be the best clue to the issues worth such further investigation). (465)

Elicitation, statistics, and corpus studies all receive a share of credit for obeying the principle of "total accountability" that, Quirk argues, is the centre of corpus linguistics. Only introspection and methodological laziness in not exploiting a corpus fully ("A corpus is not worth having unless we see everything in it") come in for criticism. His attitude to statistics is telling. Although he or his collaborator uses statistical significance to evaluate their elicitation experiments – "complementations with *-ing* versus the infinitive" – Quirk seems a reluctant fellow traveller: "I am wary of figures, coming of a Celtic race that is capable of statistical statements like 'People are dying now that never died before'" (466).

All five papers lacking a commentator, especially Quirk's, exhibit toleration of and support for work that differs markedly from their own preference, though kindest toward work by younger colleagues. Francis made the first computerized corpus but writes to acknowledge the work of those who worked manually before him. Fillmore and Halliday could not be more unlike one another in daily work, but each respects the other. Fillmore reaches out to the statistician's corpus, and Halliday brings language back to the free individual with choices to make.

The fourteen essays and responses amply reward a close reading. They discuss theoretical issues such as cognitive constraints on the individual's use of language (Wallace Chafe), and a probabilistic theory of texts (Geoffrey Leech). Others present matters of corpus design, for the International Corpus of English (Sidney Greenbaum), the Helsinki Corpus (Matti Rissanen), and Swedish corpora (Martin Gellerstam), as well as emerging standards in tagging speech (Jane A. Edwards) and encoding grammar (Geoffrey Sampson). Two papers describe the development of software for corpus analysis (John Sinclair on a system of partial but automatic analytic programs for huge corpora, and Henry Kučera on spelling and grammar checkers). Five more essays focus on applying already-available software to corpora: statistical programs to study anaphora (Douglas Biber), to parse texts (Geoffrey Sampson), and to detect rationality in mother-child conversations (Ruqaiya Hasan), and more general tools to assist the writing of the Swedish Academy grammar (Steffan Hellberg) and the teaching of languages (Graeme Kennedy).

Fourteen commentators made considered reviews of each of these papers. Many evaluations were candid. Bengt Altenberg and Göran Kjellmer, respectively, give the highest praise to Douglas Biber ("illuminating", "impressive" and "rewarding") and Graeme Kennedy ("an excellent historical survey"). I entirely agree with these judgments. At 79 pages, these two papers are very impressive research contributions.

In "Using computer-based text corpora to analyze the referential strategies of spoken and written texts", Douglas Biber employs 11,600 words in 58 text samples from the LOB and London-Lund corpora to study frequency distributions of, distance measures related to, and types of anaphoric references, an interest he shares with Wallace Chafe. Biber wrote two programs for his work. The first identified and classified the referring nouns and pronouns and established the referential chain – a useful term – to which any repeated item belonged, that is, the numbered sequence of multiple anaphors that refer back to a single referent. Once Biber manually edited this output, his second program computed the frequency count and distance measures. After analysis with SAS, a statistical system (in particular, a General Linear Models procedure), which compared results of each text type with every other text type, Biber presents the comparisons, and referential dimensions derived from them, in ten tables and five figures. They reveal many intriguing patterns, of which I can mention only a few. Spoken texts, like broadcasts, have more total anaphors or referring expressions than written texts (e.g., fiction), although conversations have fewer different referents. In contrast, spot news has the most different referents of all. Humanities academic prose has a very high proportion of "deadend" referents (ones mentioned only once) and very short chains – we might have guessed – unlike conversation, which has the fewest deadend referents and the longest chains. After factor analysis of this data, Biber associates certain features of anaphoric reference with textual dimensions he derived in 1988 "from the co-occurrence patterns among 67 surface linguistic features" and considers whether the referring expressions have dimensions of their own. They appear to have four. For instance, the first dimension Biber names "Involved referential strategies". It is characterized positively by five features – his original first dimension ("involved production"), exophoric pronouns (which refer to someone or something involved in present communication, e.g., *I*, *me*, etc.), vague pronouns (these have no specific referent in the text), average chain length (number of anaphors), and maximum distance among them – and negatively by one feature, repetition anaphors (lexical repetitions of nouns in a chain).

Conversations very often have this first dimension, but very seldom does any kind of expository prose. Anyone comparing Biber's Table 14 and Figure 6 will discover much to admire, and much to stimulate further research, for he describes the results as preliminary.

Graeme Kennedy rightly says that good language teaching focuses selectively on "Preferred ways of putting things" (the opening words of his essay title) and that finding out what those ways are asks the language teacher to pay close attention to statistical analyses of corpora. Kennedy thus identifies a critical direction for corpus studies, language education, a view shared by symposium participants Magnus Ljung, Jan Svartvik, M. A. K. Halliday, John Sinclair, and Göran Kjellmer, whose work he cites. Kennedy's richly detailed essay begins by describing a 30-year research programme by a number of linguists on English vocabulary that culminated in Michael West's *General Service List of English Words* (1953). This led teachers to focus on high-frequency words rather than the unusual. The essay then turns to implications for language teaching in corpus research since the 1960s. Kennedy rehearses research on verbs by Akira Ota, H. V. George, Martin Joos, Jennifer Coates, Janet Holmes, and others to the effect that "Most English verb forms are not used frequently enough to warrant pedagogical attention in the early stages at least" (348). Syntactic and semantic studies, and developmental research on first language acquisition, come next in Kennedy's incisive survey, which is all the same too substantial for summary here. He concludes his essay with a persuasive account of why language teaching has arrived at a stage when it can once again benefit from corpus research, of what forms that benefit will take, and of how corpus linguistics must reform itself from within so that the teaching community ceases to ignore it. This final topic has special bearing on future research. Kennedy mentions the need for corpora like ICE that cover regional varieties and registers, for previous research to be redone on larger, more reliable corpora, for systematic non-trivial studies, for "clear and transparent summaries" of corpus research in manuals written for teachers, and finally for "laborious hands-on work, particularly on semantic issues", to identify language features that are countable. This adds up to a corpus research agenda for second-language teaching and learning that would match the corpus development now underway in the ICE project. Kennedy's paper ends on this challenge.

Three other substantial papers receive polite but hard criticism from their commentators. First, Bengt Sigurd says that Geoffrey Leech does not deliver on the promise of his title, "Corpora and theories of linguistic

performance", to produce a new philosophical theory of language opposed to Chomsky's: "A more proper title might have been 'Corpus linguistics and probabilistic theories of texts". I think this topic has been nicely covered.' Sigurd has a good point and as far as I can see has no axe to grind in making it. Leech breaks down corpus-based research into three helpful paradigms, informal concordance-based, log-linear modelling for linguistics categories, and language-modelling using Markov models. This admirable discussion (pp. 113-20) should be read with Halliday's seven ways to interrogate a corpus statistically.

Second, Fred Karlsson vigorously objects to John Sinclair's essay, "The automatic analysis of corpora", as championing software tools that would put high-quality standards at risk and that abandons "perfect analysis" as a goal. Sinclair takes issue with client-funded software for specific purposes, such as language-understanding or machine-translation programs (the latter having "a succession of unfortunate results"), most of which have some specific model in mind. He argues that "we [instead] devise methods of analysis that prioritise information about the language that we can derive from the corpus" (381). His six guidelines for software design specify unlimited text size, real-time automatic operation (without any manual intervention) "at more than one level of discrimination, so as to bypass doubtful decisions", robustness, and speed. These principles favour what Sinclair calls partial parsers, each doing one well-defined task well. These include word-class tagger, collocator, lexical parser, lemmatiser, phrase finder, compounder, disambiguator, exemplifier, classifier, and typologiser. It is to be hoped that Sinclair's plea for funding of this modularized toolkit will persuade the research agencies and industrial clients to change their mind.

Third, although Benny Brodda credits the openness of attitude in Geoffrey Sampson's essay "Probabilistic parsing" and praises his willingness to develop "reusable syntactic analyses", Brodda all the same pins him to the wall on a failure to give results – "What he tries to do (and also manages to do, he claims – we have not seen a printout from an actual run, nor an actual demonstration)" – and argues that he "expresses a widely held misconception about 'productions"'. Because both Karlsson and Brodda are placed in the position of defending their own very successful but different work against researchers who explicitly reject their approaches on principle, I think their criticisms of Sinclair and Sampson are understandable, fair play. In another world, of course, these two essayists might too have been awarded immunity from commentary, or given reviewers who had comparable ideas, and so they

might have received (implicitly) higher marks. (Consider, for example, the implications of asking an introspectionist to review Randolph Quirk's paper.) Sampson describes how generative grammars fail to cope with grammatical diversity, with so-called "performance deviations" like speech repairs, and with syntactical "rules" that everyone breaks. In this way he justifies trying a different methodology, such as appears in his APRIL system (parsing by stochastic optimization). His rationale for the SUSANNE corpus is also persuasive: "taxonomic research in the grammatical domain that should yield something akin to the Linnaean taxonomy for the biological world" (437). The ensuing six-page account of the controversy that his work sparked among fellow British researchers is of less interest.

The other nine papers receive more moderate grades when they are graded. Often reviewers deliver nicely balanced assessments, recognizing limitations constructively as strengths, adducing valuable insights, suggesting extensions in method, and drawing out important implications of the essayist's work for corpus linguistics at large. Having some sympathy with the task they faced, I will mention several examples. In my opinion, the commentators are more important than the 6:1 proportion of essay pages to comments pages (311:52).

Christian Mair astutely places introspectionist Wallace Chafe, who says that "inventions without corpora are fatally limiting" (89), at some remove from the fray. In his essay "The importance of corpus linguistics to understanding the nature of language", Chafe "does not regard statistical tabulation of the corpus evidence as an end in itself but merely as a starting point for the further, qualitative analysis of those data which are interesting" (99). Yet Chafe is said to have his most original insights when he reads "the statistically insignificant residue in his data". Chafe discusses the two cognitive constraints in processing language, the "light subject constraint" and the "one new idea constraint", and observes two exceptions to the latter rule: one "in which the verb has low content" and the other "in which the entire verb-object phrase has been lexicalized". Mair himself then illustrates qualitative analysis, implicitly supporting Chafe, by looking at the use of *likely* and *probable* in the corpora.

Stig Johansson's remarks on Staffan Hellberg's essay, "Using corpus data in the Swedish Academy grammar", an account of how this project employs corpora, are also exemplary. Without a great deal of enthusiasm, Hellberg says that corpora provide authentic examples of a grammatical usage (though they mainly "represent neutral or normal written style" and seldom include rare constructions) and enable us "to test our linguistic

intuition". Understandably, Hellberg views corpora as only one means to a different and more important end. Johansson affirms Hellberg's choices (rather more warmly, however) but also reminds him that he is innovating by applying to grammars certain methods that have worked well in making dictionaries, and adds that he should go further by citing referenced examples and by employing his corpora for "hypothesis-generating, i.e. where studies of corpus material give rise to new ideas about some grammatical point" (332-33). This is just the kind of respectful helpfulness we have come to associate with Johansson.

Magnus Ljung praises Henry Kučera for his tireless efforts to translate corpus linguistics research for the use of software companies that engineer word-processing systems for the world at large. Kučera's essay, "The odd couple: The linguist and the software engineer. The struggle for high quality computerized language aids", attacks companies like WordPerfect Corporation for failing to ensure that their databases, and the algorithms for analyzing them, incorporate basic linguistic knowledge. For instance, he shows that, as lists of English spellings grow larger than 60,000 items, the verifiers employing them increasingly fail to recognize errors called collisions, where one acceptable English word form appears in place of another acceptable word. Popular spelling checkers sometimes ignore case and punctuation and often suggest dozens of corrections for unrecognized short words, with abysmal results. Kučera then describes commercial grammar correctors, especially his own published algorithm (employed in *Correct Grammar*), and stresses both their modest success and their major defects (sanctimonious prescriptive rules about the passive, the unmet challenge of highly inflected languages, and closed compounds in German and Scandinavian tongues). Ljung wonders whether these and other fundamental problems are so serious that learning aids should be set aside. He cites the inability of syntactic rules to correct some collisions (e.g., *from* and *form*), the failure of spelling checkers to treat borrowings from other languages, and the heavy prescriptivity of commercial systems, threatening to "reduce all prose produced on word processors to a kind of Newspeak unsuspected even by Orwell" (423). Candidly, Ljung observes that if corpus linguistics cannot penetrate this market, its importance will be compromised.

Martin Gellerstam begins his discussion, "Modern Swedish text corpora", with an admission that he does not exactly know what a corpus is but that it has texts by a mixed authorship that are "assembled in a predefined way ... to construct a sample of a given language" (149). The limitations of the well-defined early corpora – they could not serve many uses –

led, he argues reasonably, to a recent "text bank model", which is much larger and more diverse in texts. On this basis Gellerstam divides some 18 existing Swedish "text" corpora into two classes, one for general and the other for specific purposes. Gunnel Engwall, his commentator, then draws on her research into modern French corpora to reclassify the corpora in Gellerstam's list into two categories and six subcategories: written (literary works, learned works, newspapers, and letters) and spoken (monologue and dialogue). She also proposes that corpora should be regarded as closed sets of texts, and text banks as open sets out of which such corpora may be built. Her review casts Swedish corpora as having a more premeditated structure than does Gellerstam, who amusingly begins by saying they "may have been derivative, or 'in a sack before they got into a bag' to approximate a Swedish saying" (149).

Two papers discuss children's speech. In "Design principles in the transcription of spoken discourse", Jane Edwards bases her "minimalist standard for child language transcription", published in 1989, on seven principles of visual display for maximally readable transcription conventions, and on several matters relating to interpretability, including the normalization of variant spellings (e.g., by a conversion table) and the separate encoding of all categories rather than the use of tags that refer to multiple categories at once. Gösta Bruce, commenting on Edwards' work, states that it is, "by and large, convincing and hard to disagree with", but suggests that issues of manageability for the transcriber, and learnability, might also contribute to such a standard. Basing his remarks on his work for the IPA, Bruce then urges that speech corpora use the IPA symbol set in transcribing speech and doubts whether any "theory-neutral standard" such as Edwards suggests is possible, especially if it covers discourse structure as well as prosody.

At 51 pages, Ruqaiya Hasan's paper on measuring rationality in 22,000 messages selected from 100 hours of mother-child conversations is the longest in the collection. "Rationality in everyday talk: from process to system" would have benefited from shortening. She subdivides reasoning into the tautological and the grounded (in experience), the latter into social and logical, and social in turn into additional subcategories, including conventional and coercive. After analyzing semantically specific cases of reasoning and processing them with principal components analysis, she obtained results showing that the social status of the mother co-varied with the kind of reasoning she used. The procedural steps in her automatic processing were not clear to me, but the results were: mothers described as having a "higher autonomy profession" used logical

reasoning, but those having a "lower autonomy profession" used social reasoning. This essay tells us more about motherhood, and less about corpus work, than we might expect in a volume of this kind. Donald Hindle describes it fairly "as a finely detailed analysis ... with much inferred from the text where it is not overtly represented". As a natural-language programmer in the private sector with a hand in parsing systems like *Fidditch*, Hindle has little option but to say that "No automatic analysis of large corpora can hope to achieve the kind of detailed analysis Hasan presents" (308). Yet Hindle indicates, in two brief paragraphs, that he has personally extracted, automatically, from a 44-million-word corpus the subjects, verbs, and objects of clauses and that he has queried the resulting data-table to answer the semantic question, "What can be caused?" He has discovered that "Reasons for good things are typically not given", a result that tallies with Hasan's data and confirms that corpora can now yield primitive information about "social and semantic grounding". I would have liked to read more about Hindle's research.

Sidney Greenbaum's far shorter paper, "A new corpus of English: ICE", at 9 pages, discusses the aims and organization of a far larger enterprise, the International Corpus of English, which is interestingly restricted to texts from adults, persons of 18 years and over, but encompasses million-word sub-corpora from up to 15 countries, from Australia to Zambia, from the years 1990-93. It is instructive to compare ICE with the BNC. Greenbaum, his colleagues abroad, and his collaborator Jan Aarts at Nijmegen, are undertaking an astonishing breadth of tasks: selecting text samples by their inclusion of a wide variety of textual and social variables, inputting them, tagging them for word-class, parsing them, preparing standard tagsets and manuals, and developing software for retrieval and analysis. Jan Aarts' comments on this paper are those of a collaborator and expand on the procedures to be used by the Nijmegen TOSCA team for the tagging and parsing. Any one of their joint tasks is very difficult, but altogether they exceed in scope what the BNC evidently has in mind (Randolph Quirk indicates that its corpus will receive word-class tags, but not the kind of tags resulting from parsing), although a powerful consortium of publishers, libraries, and universities have assembled to do BNC tasks. Any comparison of the two projects testifies to the intense dedication of the ICE collaborators, and to the importance of its example to individual members of ICAME.

The last essay to be discussed, "The diachronic corpus as a window to the history of English", is the one closest to my own work. Matti

Rissanen discusses the Helsinki Corpus, a diachronic corpus of English from the eighth century to 1800, much smaller than ICE, with 400 samples of text, amounting to 1.5 million words, and as yet untagged. Yet Rissanen and his collaborators have succeeded in doing something original, although they too modestly regard their work as "only a limited and biased picture of the reality of language" (202) and state that text corpora should never be allowed to "alienate" young scholars from "the study and love of the original texts". Theirs is the first historical corpus of English, and evidently the first to be encoded with COCOA-style tags that give essential information about the author and the work, such as the type of text, and the age, gender, and social standing of the author. They are also the first to delineate the structural features of a diachronic corpus. Texts must represent adequately chronological periods of a century for Old and Middle English, and of seventy to eighty years for Late Middle and Early Modern English. As well, samples come from regional dialects found in each period (nine such dialects appear in Old and Early Middle English), reflect the writing of both sexes of "different age groups, social backgrounds and levels of education" (from Middle English on), and encompass many varying genres and types of text. Defining text types heuristically by "subject matter, purpose, discourse situation and relations between the writer and the receiver" (194), Helsinki exposes the rich, buried hoards of English in letters, state trials, and other materials found in the treasure-filled British local and national record offices. Rissanen also provides intriguing applications of his corpus: the gradual shortening of forms of (*n*)*aught* from 850 to 1250, the increase of the progressive form *be* + *-ing* from 1640 to 1710 even as periphrastic *do* decreased, and the distribution of personal pronouns across text types for all periods. The last brings to light interesting facts like Wyclif's low use of the first person plural in his homilies, as against its high frequency in the *Northern Homily Cycle*. As valuable as these insights are to historians of English, Gunnel Tottie, Rissanen's commentator, suggests not only that they will be important to experts in modern English too but that a comparable diachronic corpus be made for the post-1800 periods so as to facilitate comparisons. She justifies this need from her own work in tracing the distribution of indefinite determiners in non-assertive clauses.

Let me close my extended review with some personal opinions about corpus linguistics as shown in this collection.

- The accomplishments of corpus work, past (such as relate to Brown, LOB, London-Lund, and the Swedish corpora) and present (Helsinki), and its new projects underway (ICE and BNC), give ICAME just cause to be proud and confident. It need not worry about the indifference of Noam Chomsky and his disciples, the ignorance of language teachers, and the feast-or-famine attentions of private-sector clients. The field has proved that a representative language corpus, closed for the purpose of exhaustive tagging, parsing, and analysis, is an essential scientific tool, central to all types of linguistic research and to most practical applications associated with it. As Sinclair and Kučera have proved before and continue to display in their solid work, any language industry undertaking text-processing software development is uncompetitive without a team of corpus linguists at their side.
- Defensiveness can lead to closures of less welcome kinds. Corpus linguistics should encourage innovative linguistic research on open text banks as much as it does on closed corpora. I especially regret finding no essay on the monitor corpus. Given the astounding recent growth of electronic libraries and data banks on the Internet, corpus linguistics is already awash with more data than it can handle. Existing research on how to derive linguistic information, both diachronic and synchronic, from an open and ever-increasing ocean of text will be critical to the ability of corpus linguistics to develop in the 1990s as it has in the past decades.
- Randolph Quirk and Graeme Kennedy indicate other large fields that would benefit from corpus linguistics expertise: research on, and teaching of, non-English and second languages. Literary history and perhaps philosophy, both text-dominated fields, also have many untapped uses for corpora.
- Halliday, Leech, Biber, and Sampson make a very persuasive case that corpus linguists should use statistical tests in their analyses of corpora and understand statistical models of how languages work. The courses, manuals, and textbooks needed for these purposes are not yet available and have some priority.
- Finding statistical significance in the distribution and co-occurrence of language features, as among types of writing, does not help us understand the significance of those patterns in human terms. Svartvik, Fillmore, and Chafe all remind us that we are studying the mind at work. Current *experimental* research in cognitive psychology, mainly through *elicitation* but occasionally with small corpora, is gradually

leading us – with the help of *statistical* tests – to an understanding of the constraints within which the human brain speaks, listens, reads, and writes. These constraints explain the patterns that corpus studies are discovering in texts.

Nelleke Oostdijk and **Pieter de Haan** (eds.). *Corpus-based research into language. In honour of Jan Aarts*. Amsterdam and Atlanta, GA: Rodopi, 1994. 279 pp. ISBN: 90-5183-588-4. Reviewed by **Udo Fries**, Universität Zürich, Switzerland.

This is a *festschrift* (though the term is avoided in the book) in honour of Jan Aarts on his 60th birthday. Nelleke Ooostdijk and Pieter de Haan have given us a stimulating collection of 15 papers (including their own *Introduction*, in which they survey the field of corpus linguistics).

To begin with, Flor Aarts has written a delightful little masterpiece about Jan Aarts, which everybody who is curious about the relationship between the two Aartses should not fail to read. The remainder of the book is divided into three slightly unequal topical sections, Part I: The encoding and tagging of corpora (5 papers), Part II: Parsing and databases (6 papers), Part III: Linguistic exploration of the data (3 papers), followed by a reference section and a list of contributors. Authors and editors have done their best to unify this collection of very different papers by referring to them as chapters, by introducing occasional cross-references, and by the common bibliography, which is a very useful contribution in its own right – and avoids unnecessary and boring duplication.

It is difficult to describe and define the ideal reader for this volume. Some of the papers, especially in the first section, are clearly aimed at the uninitiated in corpus linguistics or certain areas of it and provide a useful introduction to the world of English corpora; others presuppose a great deal of expert knowledge of the more technical aspects of corpus design, while the last group of papers will be of interest not only to the corpus-linguist but to the student and scholar of Modern English in general.

In Part I, Stig Johansson (*Continuity and change in the encoding of computer corpora*, 13–31) addresses the beginner and the future corpus compiler, who are provided with a description of the tagged version of

the LOB corpus, a comparison between earlier encoding systems and those used today under the influence of TEI, followed by an introduction to the TEI guidelines. Sidney Greenbaum and Ni Yibin (*Tagging the British ICE Corpus: English word classes*, 33–45) also have the beginner in mind when they compare the tagging system CLAWS1 for LOB and its development for the ICE Corpus. They present an outline of the targets of the ICE corpora, which are useful for grammatical analysis rather than lexical studies, and they argue for degrees of tagsets (reduced tagsets) for different purposes. Geoffrey Leech, Roger Garside, and Michael Bryant (*The large-scale grammatical tagging of text: Experience with the British National Corpus*, 47–63) address the potential users of the BNC, making them aware of the problems involved in large-scale tagging (with CLAWS 4), where tags can no longer be manually corrected, but various means of improving on automatic tagging must be used, and the result is no longer the 100% "correct" corpus. Instead, users will get a useful tool they can work with for their specific purposes. Both Greenbaum and Leech conclude their contributions with a list of tags for ICE (p.36) and CLAWS: C5 (p.62–63) respectively, with Greenbaum also telling the reader where to get the complete set. Leech et al. tend to take this type of information for granted.

Yet another type of corpus is the topic of Willem Meijs' contribution (*Computerized lexicons and theoretical models*, 65–78), which deals with the LDOCE in its machine-readable (MRD) form and gives a survey of what the Amsterdam group has done with it and its relation to the Nijmegen TOSCA project. The addressee is most likely the tagging and parsing corpus linguist. It becomes clear that, in so wide a field, not everybody is equally aware of the other's work. This becomes apparent, when one reads the final chapter of the first section by Louise Guthrie, Joe Guthrie, and Jim Cowie (*Resolving lexical ambiguity*, 79–93), in which there is no link to the previous paper by Meijs (and vice versa). The studies reported in Guthrie et al. seem to be more general and more ambitious, whereas Meijs' approach is more down to earth. The example of *bridge* in the Wall Street Journal is a very concrete and illuminating example in Guthrie's paper.

None of the papers in the first section is so specialised that it would be of interest only to the select few who are in the forefront of research. Aimed rather at the general reader with some knowledge, these papers show extremely well the current developments in the tagging and encoding of computerized corpora, and what is happening on both sides of the Atlantic.

The second part presupposes a good knowledge of grammars and parsing. Ted Briscoe (*Prospects for practical parsing of unrestricted text: Robust statistical parsing techniques*, 97–119) talks about experiments with robust parsing techniques, which have become possible because of the increasing availability of genuinely wide-coverage grammars couched in computationally tractable formalisms (such as the TOSCA and ANLT-grammars). Fred Karlsson's contribution (*Robust parsing of unconstrained text*, 121–142) shows more didactic qualities. While the uninitiated will be pretty much at a loss in Briscoe's paper, they will catch up again in Karlsson's chapter, which makes it clear what *robust parsing*, what a *Constraint Grammar* and what CG syntax are, and how one achieves results in these areas; and, even more important, where to turn for a test run of ENGCG (by e-mailing to Helsinki). ENGCG claims to be more successful than CLAWS1 and PARTS, and will be used for the 200-million-word corpus of COBUILD. Karlsson's contribution is a model of how a difficult subject can be presented in an easily readable way. Incidentally, he presents the state of the art by September 1993; the book appeared only a few months later, which is important for a publication of this kind, but by no means common practice.

The chapter by Clive Souter and Eric Atwell (*Using parsed corpora: A review of current practice*, 143–158) is a very reader-friendly survey of parsed corpora (including the addresses of where to obtain these corpora) and the types of parsers available, answering the question of what a parsed corpus looks like (labelled brackets or numbers), and presenting as one of its conclusions the disillusioning acknowledgement that a parsed corpus is not the answer or solution to all problems. Ezra Black (*An experiment in customizing the Lancaster Treebank*, 159–168) presents an analysis of the determiner phrase, the adverb phrase, and compound nominal expressions in order to improve parsers. This is a report about a very specific problem; for the general reader, it gives an impression of the type of thought given to such problems – and, perhaps, a reminder of the complexity of language structures.

By the time readers arrive at Geoffrey Sampson's contribution (*SUSANNE: A Domesday Book of English grammar*, 169–187) they will have met SUSANNE several times. Now they get a detailed introduction to it and all the information necessary for a retrieval of a copy of this corpus.

Part II concludes with William Gale and Kenneth Church (*What is wrong with adding one?* 189–198), who present a very specific statistical

problem which occurs with corpora that are not big enough to include all the items you may want to investigate. In this case, the question is what to do if there is not a single occurrence of an item in the corpus. This is an exposé for experts in stastistics and mathematics.

Part III begins with a study at least touching on an area of corpus linguistics that is not represented in this volume: diachronic corpora. Douglas Biber and Edward Finegan (*Intra-textual variation within medical research articles*, 201–221) analyse part of their new ARCHER corpus. The medical sub-corpus contains medical articles from the *New England Journal of Medicine* and the *Scottish Medical Journal*. Altogether, 19 articles, all of which show an I-M-R-D-Structure (Introduction, Method, Results, Discussion) and all from 1985, are compared to each other, but also, and more importantly, to an overall reference corpus. The individual sections of medical articles are situated among other genres in the multi-dimensional analysis of English developed in Biber (1988). The article can, indeed, be read as a very useful, if brief presentation of Biber (1988), but also as a study on differences between British English and American English written registers. A survey of the diachronic dimension of the ARCHER Corpus should show the evolution of these registers during the last three centuries.

Bengt Altenberg's study (*On the functions of* such *in spoken and written English.*, 221–240) can be regarded as a perfect example of how to make the best use of computerized corpora. He proposes his own theory – based on previous treatments of the problem, gives a wide variety of examples – taken from the vast number of occurrences in the corpora, and analyses the stylistical distribution in different genres. This is a theoretical study of a notoriously difficult problem of English syntax and semantics – which goes well beyond previous studies and sets a new standard for the treatment of *such*. It makes full use of the possibilities of a corpus: providing vast numbers of examples that would not necessarily occur to an armchair linguist (or which could be more easily discarded), it provides useful insights into their distribution over various text genres, and, last but not least, shows the limitations and possibilities of future research both with synchronic and, more importantly, with diachronic corpora.

The volume closes with a study by Anna-Brita Stenström and Jan Svartvik (*Imparsable speech: Repeats and other nonfluencies in spoken English*, 241–254). The authors take as their starting point problems that occur with the parsing of the ICE Corpus. They establish a typology of nonfluency in speech with special emphasis on pronoun repeats.

Taking their data from different sets of the London-Lund Corpus, they are able to offer corpus-specific findings, which show clear differences compared to previous research, but also differences between individual text-types (ranging from court examination and proceedings to multi-party chats). The scale of nonfluency which they establish will be the basis of future research.

The three articles in the last section will assure this volume[1] a more permanent relevance, at a time in the future when the problems of tagging and parsing corpora will have been solved. But this is still a long way off.

Note

1 The book is very well produced, with only a few minor errors. In the table of contents and in the headers of the first section, this part of the book is called *The encoding and tagging of corpora*, but on the title page of Part I, p. 11, *The tagging and encoding of corpora*. In Meijs' paper, p. 70, the reference to Akkerman *et al*., 1985 leads to no entry in the bibliography at the end of the volume. On p. 73, 2nd paragraph, in the last line but one, read *subject field hierarchy* instead of *box code hierarchy*. In the chapter by Guthrie et al., p. 80, the reference to Gale (1992) is not listed in the bibliography; it may refer to Gale and Yarowsky (1992). Similarly, in Briscoe's paper, p. 118, a reference to Wu (1992) leads at best to Wu (1990), and in the contribution by Gale and Church, p. 190, a reference to Church (1989) may refer to Church (1988), a reference which, incidentally, gives *Ausin, Texas* as the place of publication (p. 260)! The most deplorable misprint occurs on the very last page of the text (second word on p. 252): a reference to *distant methatheses*. Brush up your Greek: μεταθεσις should be rendered as *metathesis*.

Reference

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

Udo Fries, **Gunnel Tottie** and **Peter Schneider** (eds). *Creating and using English language corpora*. Language and Computers: Studies in Practical Linguistics, 13, 1994. Amsterdam Atlanta, GA: Rodopi. iii + 203 pages. ISBN: 90-5183-629-5. Reviewed by **Henk Barkema**, University of Nijmegen.

The volume *Creating and Using English Language Corpora* consists of proceedings from the XIVth ICAME conference on English language research on computerized corpora, which was held in Switzerland in May 1993. It gives an accurate state-of-the-art impression of work nowadays going on within the several fields of corpus linguistics.

The portrait of the era which it provides is perhaps slightly out of balance, as one strand of activity is somewhat underrepresented, namely that of automatic corpus annotation. Only two chapters (one by Nancy Belmore and another by Atro Voutilainen and Juha Heikkilä) deal with this topic. However, other volumes on the same *Language and Computers* shelf make up for this imbalance.

Let me give a thematic inventory of *Creating and Using English Language Corpora*. One part of the book consists of descriptive studies – a distinction can be made here between studies of historical, diachronic and contemporary English. Some of these focus on lexical, some on lexico-grammatical and others on grammatical issues. In relation to contemporary English, we can make a distinction between contrastive and non-contrastive corpus research. Another part is about software: about how exploitation tools can be used efficiently, and how analysis tools can be improved. I will not discuss each of the seventeen chapters in the book separately: a brief overview is provided by the editors in the introduction. Instead, I would like to pick out a few bits and pieces which I found particularly interesting.

For example, in a contribution entitled 'Is *see* becoming a conjunction? The study of grammaticalisation as a meeting ground for corpus linguistics and grammatical theory' Christian Mair says two sensible things about language theoreticians: 1) corpus linguists often have to help them to land softly back on terra firma; 2) corpus linguists can benefit from ideas put forward by theoretical linguists. The first remark is illustrated in Helena Raumoulin-Brunberg's chapter 'The position of adjectival modifiers in Late Middle English noun phrases'. She uses the Helsinki corpus to convincingly refute the claim (put forward by theorists) that in Late Middle English adjectives predominantly must have taken the function of noun phrase postmodifier. By discussing the notion of

'grammaticalisation' – a topic which for some time has been popular with language typologists – Mair himself illustrates his second remark.

New corpora sometimes open the way to new, exciting research questions. An example is ARCHER, an acronym of 'A Representative Corpus of Historical English Registers'. This 1.7 million-word corpus of American and British English, compiled at the universities of Northern Arizona and Southern California under the supervision of Douglas Biber and Edward Finegan, nicely bridges the gap that for some time existed between the Helsinki corpus (Old to Late Modern English) on the one hand and the first present-day English corpora dating from the early sixties, such as LOB, Brown and London-Lund, on the other. As Biber and Finegan, together with Dwight Atkinson describe in 'ARCHER and its challenges: compiling and exploring a representative corpus of historical English registers' (and illustrate in a typical Biber-and-Fineganian fashion), the corpus can be exploited in a variety of (synchronic, diachronic and contrastive) ways; by means of advanced statistical techniques they arrive at accessible and intuitively natural descriptions of texts.

Another example of a new type is the parallel corpus; in 'Towards an English-Norwegian parallel corpus' Stig Johansson and Knut Hofland remark that the study of bilingual and multilingual corpora is still in its infancy. With their corpus they will be able to make up for this. It will be used for various new types of contrastive study, as well as for the examination of translation problems and a phenomenon they call 'translationese': deviant language use that is the result of translation.

The research reported on by Jan Svartvik, Olof Ekedahl and Bryan Mosey in 'Public Speaking' is of special importance for the increasing number of linguists who are interested in transcribed spoken English and who want to know how they should split their texts up into prosodic chunks. As part of their Public Speaking project, Svartvik and his team try to discover which segmentation speakers use to divide their texts into tone units.

Improvement of existing software is the concern of a number of contributors. In 'Towards a grammar checker for learners of English' Sylviane Granger and Fanny Meunier discuss the criteria which such a tool should meet in order to assist language learners to produce texts without grammatical mistakes. They put three programs to the test and come to the refreshing conclusion that producers of grammar checkers should consult EFL/ESL specialists to find out what language learners really need. At the same time Nancy Belmore's concern in a chapter

poetically entitled 'Contrasting the Brown corpus as tagged at Brown with the Brown corpus as tagged by CLAWS1' is the improvement of the quality of grammatical analysis tools. By means of a relational database, she compares the Brown and CLAWS1 taggers, which make use of the same tag set. By studying the contexts in which both taggers fail, she tries to establish how the quality of such tools can be improved.

In the last (but by no means the least) chapter of the book Atro Voutilainen and Juha Heikkilä give a description of 'An English Constraint Grammar (ENGCG): a surface-syntactic parser of English'. Judging from their assessment, the system must be extremely fast (a quick calculation tells me that it can process a 200 million-word corpus in less than a week (provided one is in the possession of the right hardware), with 94.5% of the wordclass tags correct and unambiguous. This must be the lexicographer's dream come true, who, until recently, nearly seemed to drown in massive pools of raw corpus data. The syntactic component of the parser has its pros and cons. In relation to giga-corpora, its advantage is that it blindly labels no less than 80% of all words with unambiguous and correct syntactic tags: a score which will be improved as soon as more constraints have been added. The price for the tool's efficiency is that it only assigns syntactic function labels to individual words, while of modifying words it only indicates in which direction (to the left or to the right) the heads can be found – something which owners of large corpora (who are predominantly interested in lexicographical or lexico-grammatical issues) will be happy to accept. It therefore fills a lacuna, left open by the much more labour-intensive rankscale constituent parsers, which are better-suited for the analysis of much smaller corpora that can be used for purely syntactic research.

While reading the book, I noticed two things I do not quite understand. The first is why relatively many linguists still carry out grammatical or lexico-grammatical research on the basis of entirely raw corpora, which is surprising in view of the fact that nowadays a great many efficient taggers are available (three of which are mentioned in this book), while a number of skeleton, automatic and interactive parsers have been around for some time. What I find even more surprising, is that no mention whatsoever is made of lemma-tagging. The addition of such tags to a corpus tagged with wordclass labels must be a relatively easy enterprise and would save linguists with an interest in lexico-grammatical issues a lot of tedious work.

To conclude: for those of you who want to know more about the articles discussed in this review, about East African or Hong Kong

English corpora, about the influence of American and British English on Australian verb inflections or the development of English adverb forms throughout the ages, about statistical techniques to examine the fixedness of recurrent word combinations, the grammar of lexicalised expressions or text styles, or want to know how a dubious method used in British courts has been exposed by corpus linguists purely on theoretical grounds, there's only one option: buy this Swiss timepiece, and read it!

Dieter Mindt. *Zeitbezug im Englischen: Eine didaktische Grammatik des englischen Futurs*. Tübinger Beiträge zur Linguistik 372. Tübingen: Gunter Narr, 1992. 328 pp. ISBN 3-8233-4227-4. Reviewed by **Herman Wekker**, University of Groningen.

In 1989 I wrote a review for the *ICAME Journal* (vol.13, pp. 81-83) of Dieter Mindt's previous book entitled *Sprache, Grammatik, Unterrichtsgrammatik: Futurischer Zeitbezug im Englischen* and published in 1987. The resounding message of that book was that corpus studies should be applied to the improvement of language teaching materials. I noted then that Mindt's work had a great deal to offer to textbook writers, teachers and teaching methodologists because it is immediately relevant to the practical needs of teachers and learners of English as a foreign language. His research goal over the years has been to find a new way of compiling pedagogical grammars by using an electronic database for linguistic analysis. The area that he and his team at the Free University in Berlin have focused on since 1979 is that of future time reference in present-day British English. Their ultimate aim was to arrive at a (plan for) pedagogical grammar of futurity in English. The project consisted in a detailed comparison of information on eight expressions of futurity found in a large corpus of English and the way futurity is treated in two widely used learners' grammars. The corpus consisted of two parts: 170,000 words of conversational texts taken from the Survey of English Usage (recorded between 1953 and 1976), and drama texts (184,000 words; published between 1971 and 1980). In addition, he examined two English coursebooks (*English H* and *Learning English Modern Course*) which are widely used in Germany, for comparison with the corpus data (about 281,000 words). In total the materials studied amounted to about 635,000 words. The eight expressions were:

will + infinitive, *shall* + infinitive, *going to* + infinitive, present progressive, simple present, *will* + progressive infinitive, *shall* + progressive infinitive, and *going to* + progressive infinitive. The results of Mindt's sophisticated analysis were interesting and sometimes quite surprising. He found that the current reference grammars of English provide insufficient and also misleading information on the expression of the future. It is indeed a miracle that our teaching materials are as good and authentic as they are. He found that there is a high degree of homogeneity in the use of future time expressions in his two subcorpora (Conversation texts and Play texts). He noticed the high overall frequency of *will* in comparison to *going to*, the unexpected importance of *shall* and the striking infrequency of the remaining expressions. In the two coursebooks which were examined he observed an over-emphasis on *going to* in relation to *will*, and the complete absence of *shall*.

The present volume by Mindt, entitled *Zeitbezug im Englischen*, marks the end of the Berlin project on futurity. We are not told whether they are planning to apply the same techniques to other areas of the grammar, as I recommended in my 1989 review of Mindt's previous book. The method used as well as the materials and the expressions analysed have remained the same as before. The new book provides not only a summary of the old results but also adds further details of the team's analytical corpus work. The additional information concerns the morphology, syntax and semantics of future time expressions in English, still with a view to the planning and design of a pedagogical grammar. Mindt repeats the distinction he makes between what he calls didactic grammars and pedagogical grammars, the latter being derived from the former. His model involves three steps: 1) compilation of a corpus for specific language teaching purposes, 2) derivation from the corpus of a didactic grammar, and 3) planning of a pedagogical grammar on the basis of the didactic grammar and of language teaching methodology (selection, grading, presentation, etc.). This seems to me a powerful model to work with, as I wrote in 1989, but I have no indication that Mindt has actually produced a didactic grammar of this kind, let alone a pedagogical one, for the expression of futurity or any other topic. His work has been mainly concerned with the analysis of the corpus. I am not aware of any plans to continue these useful explorations.

The present volume consists of six chapters. The first deals with the main principles and assumptions of the research project. The second is concerned with morphology, the third with syntax, and the fourth with semantics. Chapter 5 gives a summary of the findings with a discussion,

and chapter 6 draws conclusions from the results suggesting a perspective for further research. There is a full bibliography of works on corpus linguistics and futurity as well as a good Index. Finally, the book contains a 110-page Appendix with tables and diagrams. Like its predecessor, the book is written in German instead of English.

What is new in the book under review is not so much the approach or the basic idea, but the completeness of treatment. For the first time we now have a comprehensive analysis of the distribution, co-occurrence and shades of meaning of English future expressions on the basis of electronic data. Apart from his own corpus of conversational and drama texts (the CONV and PLAYS subcorpora), Mindt has now also made use of numerous examples of future reference quoted by previous scholars. As far as written English is concerned, he leans heavily on my 1976 dissertation on *The Expression of Future Time in Contemporary British English*. I am grateful to him for incorporating and correcting some of my own findings. Perhaps it would have been even better if he had used a new, larger and more up-to-date corpus; the texts in his corpus were at least 12 years old when the book was published.

The new book gives us more information, for example, about the frequency of *will* (64%) vs *going to* (16%); the present progressive and the simple present each occur less than 10%, in the main corpus. The other future expressions are extremely rare (apart from *shall*, which is mainly restricted to the first person sing.). In the teaching materials, *will* is clearly underrepresented, *going to* is overrepresented and future *shall* hardly occurs at all. It is striking that there are no great differences between the two subcorpora, but that there is a considerable discrepancy with the teaching materials. The cluster analysis yields interesting results about the type of main verb use, the co-occurrence with future time specifiers, the degree of contingency expressed by each of the constructions etc. From the electronic database it should be possible to derive a variety of pedagogical products for different target groups. Ultimately, this will contribute to the further improvement of English language teaching.

Mindt and his team are to be congratulated on the completion of this part of their long-term research project. It is very valuable work which they have done over the past dozen years, not only from the linguistic point of view, but also because of the pedagogical perspective their work has always adopted. It is to be hoped that this kind of educational research will continue in Berlin and elsewhere.

Anna-Brita Stenström. *An introduction to spoken interaction*. London: Longman. Learning about Language Series. 1994. pp xiv + 238. Reviewed by **Gerry Knowles**, Department of Linguistics & Modern English Language, Lancaster University, UK.

Conversation analysis is a relatively new and interdisciplinary subject which is approached in very different ways by scholars in the contributing disciplines. This can make it difficult for the beginner or the outsider to obtain a good overall picture of the field. It also means that what makes a suitable introductory text may be different for students of sociology and students of linguistics. This book presents a clear and systematic account for linguists.

Contributions to the *Learning about Language Series* are intended to be summaries for the benefit of readers without a previous knowledge of the field. In these circumstances it would be easy to put together a digest of other people's work. This book is much more than that. It brings together ideas from different sources and fashions them into a consistent model, with the parts identified, labelled and related to each other. I quickly found myself reading it for my own benefit rather than as a reviewer. I shall take for granted that the book is to be recommended highly both for the clarity of the exposition, and for the map of the field which it provides, and I shall turn my attention to its contribution to current work in corpus linguistics.

The book is informed throughout by the extensive experience of the author and her colleagues of working on the London-Lund Corpus. From the point of view of the corpus linguist, the topics raised are among those which will have to be tackled over the next few years in the annotation and analysis of interactive spoken corpora. An important question is whether conversation analysis has yet achieved the combination of theoretical rigour and practical robustness which is required to deal exhaustively and consistently with large bodies of natural data. On the evidence of the book much has already been achieved, but unsolved problems remain. In these circumstances the purpose of a critical review is to identify possible directions for future research.

From a theoretical point of view, the book moves out into new areas, and combines old and new approaches to language structure. This leads to an interesting tension between on the one hand those claims which follow deductively from conventional linguistic assumptions, and on the other hand those claims which follow from an empirical study of the data. This applies to segmentation and to categorisation.

The structure of conversation is presented in the form of a tree (p32), of a kind familiar for example in metrical phonology, in which sequences of units on one level are made up of sequences of units on the level below. Closer inspection, however, reveals that these units are not all of the same kind. Some, for example, belong to an initial position and others to a final position. The telephone conversation on p12 has opening and closing phases, and some discourse markers (p63) introduce units of discourse. In my view, this kind of structure is actually too complex to be represented by a tree, and what is required is some kind of transition network with a formal procedure for progressing from the beginning to the end of a unit.

A network would have the additional advantage of providing a more principled approach to segmentation. In the answer (p211) to the first exercise (which, incidentally, I found rather difficult) a hesitation ("erm") is deemed to complete Exchange 2 introduced by a question, whereas a follow-up question and answer in Exchange 4 are treated as part of the preceding exchange. To me this looks arbitrary. Some of the things said in conversation – asides, hesitations, backchannels, follow-ups and afterthoughts – relate in different ways to the main flow, and these can be handled by a network model. Progress through the network must also include the possibility of aborting and starting again.

The units at different levels in the tree form a hierarchy: transaction, exchange, turn, move, act. This apparently conforms to what phonologists call the *strict layer hypothesis*, according to which units consist of integral numbers of units of the level below, and units cannot straddle the boundary between higher level units. The point is explicitly made, however, that the data does not necessarily pattern in this way at all. Turns overlap when participants speak simultaneously; backchannels are not 'proper turns' (p5) and seem to be excluded from the hierarchy. Although in the case of *chaining* sequences (p51), exchange boundaries coincide with the ends of turns, in the case of *coupling* sequences (p53), the exchange boundary comes in the middle of a turn. At a lower level, when a speaker finishes off someone else's words, the turn boundary comes in the middle of a move. There are also other units – pause units, performance units, tone units and information units (pp7 – 10) – which have an ill-defined relationship not only to each other but also to the hierarchy. Some kind of theoretical modification is required here. The internal structure of units and the distribution of boundary markers must be treated as separate problems. In prototypical cases, boundary markers occur conveniently at the ends of units. The problem with real

data – here as elsewhere – is that it does not always conform to the prototype.

In some cases lists are given of units occurring at each level. Types of move are listed on p36, and acts are divided into primary (p39), secondary (p44) and complementary acts (p46). These are all introduced by the non-committal formula 'The following (units) have been identified', which leaves open the question of whether they are a complete set (like a morphological paradigm or a phoneme inventory) or a part of an open list (like the set of nouns or verbs). In fact, units relate to each other in several different ways. Taking for example primary acts, <disagree> contrasts with <agree> and is in complementary distribution with <reject> (being a negative response to a different kind of act), while <question> is complemented by <answer>, but also forms a scale with <query> and <disagree>. Acts can even instantiate each other, e.g. an <answer> can occur as an <accept>, an <evade> or a <reject> (p118). In view of the large number of categories and the complex relationships among them, it would be difficult in practice to assign a unique label to each unit in a text.

These theoretical difficulties are of course problems of the subject in general, and are not specific to this book. The corpus-based approach, which is specific to the book, is one that offers a solution. The category labels could also be used, for example, to annotate a corpus. More precisely, an attempt to apply them systematically would reveal the problems and lead to the design of an improved annotation set. Ideally, a sample of annotated text could have been included as an appendix to the book.

I would also have liked to see the labels and notation conventions used to annotate the examples cited in the text. They are used to highlight technical terms in the main text, e.g. '<alerts> do not always have the intended effect' (p74), but the <alert> referred to – *HÈY# – is marked not with angle brackets but with prosodic notation. It has to be said that the prosodic notation is not always relevant, whereas the structural information would always be helpful.

An area which might have been investigated in a book introducing *spoken interaction* is the manner in which power relationships are established and negotiated. The data reported provides a number of examples. Turns in an exchange are not of equal status, e.g. speakers who ask questions and respond to the answer with a follow-up such as *I see* (p49) are assuming the right to do so. Consider also the manner in which questions may be answered. An example (p12) is reproduced

here in orthographic notation:

> B: Mr Hurd, it's professor Clark's secretary from Paramilitary College.
> A: Oh yes?

A uses a rising tone on *yes*, which indicates that at this point he assumes a superior position. His reply would have been totally inappropriate if the caller had been his vice-chancellor. Chapter 3 deals with a range of interactional strategies – turn holding and yielding, backchannelling, initiating – as though all speakers were in unchanging relationships of equality.

Finally, is this book suitable for its intended readers? The theoretical problems which have been highlighted in this review are shared by other introductory textbooks. It is after all considered perfectly acceptable to introduce other linguistic concepts – *phoneme*, *tone group*, *adverb*, and even *word* and *sentence* – as though they were well defined. Beginners using such textbooks can be protected from the problems if they are given invented data to work on, but not if they work on corpus data. Much depends here on the skill and sensitivity of the teacher, who has to understand the problems of the bright student who has discovered the shortcomings of the system, whether the problem relates to phonemes, adverbs or conversation structure. Used in the appropriate pedagogical context, this book will be eminently suitable not only for corpus linguists, but also for beginners.

Sonia Vandepitte. *A pragmatic study of the expression and the interpretation of causality: Conjuncts and conjunctions in modern spoken British English*. Brussel: Paleis der Academiën, 1993. 209 pp. Reviewed by **Hilde Hasselgård**, University of Oslo.

This book is a revised version of the author's PhD dissertation. It aims to examine causal relations from a variety of angles, from lexical and syntactic to pragmatic and cognitive. The study is confined to those expressions of causality in which (at least) two finite clauses are connected by means of a conjunction, conjunct, or some other type of phrase with a causal meaning. The term *conjunctional* is used to cover all these types of relators. Furthermore, a distinction is made between *causal* and *consecutive* conjunctionals; respectively those that introduce

a clause expressing the cause of another state of affairs (such as *because, for, the reason ... is*), and those that introduce a clause expressing the consequence of another state of affairs (such as *so, consequently, that's why*).

The corpus for the investigation consists of texts representing four different registers: conversation (9 texts from the London-Lund Corpus), political interviews (interviews with politicians), various interviews (interviews with people other than politicians) and parliamentary oral answers. The two interview categories have been taken from *Radio 4*. 375 examples have been collected from each register. This material constitutes the basis for the quantitative part of the investigation. The study is not, however, entirely corpus-based, in that the material has been supplemented with examples from outside the corpus as well as invented examples (consistently marked as such), including some that are deliberately unacceptable.

It may be noted that in excerpting examples, Vandepitte seems to have maximized the number of causal links by consistently interpreting a link as causal in cases of (potential) ambiguity, such as in (1), where the relation may be interpreted as causal or temporal.

1 It [...] is now that he is on the backbenches that he is interested in the housing programme. (invented example, p 45)

In the same vein, Vandepitte takes a liberal view when judging the acceptability of a construction, and accepts any construction for which a context can be imagined, even if it is as unusual as "spoken in a triumphant tone" (p 124), or "pronounced parenthetically" (p 127).

Chapter II establishes the lexical inventory of causal/consecutive conjunctionals as attested in her material. The syntactic characteristics of the conjunctionals are examined within a generative framework, in order to establish whether they are syntactically equivalent. The generativist distinction between syntax and lexis is upheld, so that semantics and selectional restrictions, belonging to lexis, do not enter this part of the discussion. Applying various syntactic tests (clefting, adverbial modification, movement to another position, obligatoriness of *move alpha*) Vandepitte arrives at four sets of conjunctionals which are syntactically equivalent, though perhaps not interchangeable for pragmatic reasons (p 59). It may be noted, however, that not all the conjunctionals in the 'lexical inventory' (p 41) appear in one of the four sets, presumably because they resist grouping on the basis of the syntactic criteria.

In many ways Chapter III constitutes the main part of the book, focusing on pragmatic and cognitive aspects of causal expressions. It investigates whether syntactically equivalent conjunctionals are interchangeable, and whether some conjunctionals can be semantically and/or pragmatically equivalent. This is done by examining carefully the contexts in which causal relations are expressed and whether the context imposes any restrictions on the selection of conjunctional.

A key concept here is the speaker's *propositional attitude*, i.e. the extent to which the speaker regards a given state of affairs as true or desirable. The propositional attitude can concern the causal relation itself, or the states of affairs that are causally related. It is found, for example, that some restrictions apply as to the selection of conjunctional in cases where the conjunctional is negated; i.e. where the speaker believes that a causal relation is not a true state of affairs, such as in (2). Similar restrictions apply to the use of conjunctionals in questions.

2 He killed her *not* <u>because</u> she had betrayed him, but for some other reason. (invented example, p 67)[1]

There are, however, few examples in the corpus of a negated causal relation (7 out of 1500) and most of the examples given are constructed.

Another key concept is *the speaker's knowledge of the universe*, which pertains to knowledge about the context and about the type of causal relation to be expressed. As an example, register is shown to affect the choice of conjunction, in that the conjunctionals are unevenly distributed over the corpus texts. The category of Parliamentary oral answers seems to stand out by having much higher proportions of *as* and *since* than the other three, mainly at the cost of *because*, which is nevertheless the most frequent causal conjunctional in all the registers. As regards consecutive conjunctionals, *so* is the most frequent one, except in Parliamentary oral answers, where instead there are more instances of *so that* and *therefore* than in the other registers. A table on p 84 presents a list of the 10 most frequent conjunctionals, not unexpectedly with *because* and *so* at the top (together they account for nearly 2/3 of the total number of examples in the corpus). Since the registers in the material do not represent the same amount of text, a small-scale frequency count is carried out on 2,500 words from each register, revealing that causal relations are most frequently expressed in conversation, and that causality is probably not a characteristic of argumentative discourse.

Information structure is dealt with in terms of *manifestness*. It is

claimed that the selection of conjunctional is to some extent dependent on whether the cause or the consequence is manifest; i.e. easily retrievable for the listener. For example the three most frequent causal conjunctions *as, because, since* differ in that *because* tends to introduce a proposition which is not manifest, while *as* is often used to introduce a manifest proposition, with *since* somewhere in the middle.

Only a small set of conjunctionals can introduce an answer to a *why*-question. These are claimed to be *because, on the grounds that, that's because, the grounds are that,* and *the reason is that*. However, the corpus yields few examples of this type, and they are all introduced by *because*. The other conjunctionals are illustrated by means of invented examples. Invented examples are also used to show the unacceptability of some other conjunctionals in this position, such as (3).

3 – Why is aircraft noise a particular problem here?
– ?[2] As/Since we're close to Heathrow Airport.
(invented example, p 96)

It is hypothesized that the restrictions on the use of conjunctionals in responses to *why*-questions may be related to those that are to do with manifestness, since such responses typically provide information which is not manifest in the listener's mind.

Moreover, conjunctional selection may depend on how manifest the speaker wants each part of the causal situation to become after the utterance. For example, if it is the causal relation itself that is meant to stand out, the conjunctional will tend to be stressed. However, not all conjunctionals can be stressed, according to Vandepitte (p 102). Among these are *for, in that, hence, thus*. This type of statement is of course dangerous, because it takes only one example to falsify the claim, and indeed, (4) is an example of a stressed *thus* from a part of the LLC which is not included in Vandepitte's corpus. Similar examples were found with *hence.*

4 to ^all m\urderers# the ^Homicide Act of :nineteen fifty-:seven of course di"!v\ided# - [?@:] ^sentences be:tween - !capital p/unishment# - and . "n\on-capital {p\unishment#}# - **"th\us#** - [@:m] . for ex\ample# - a ^man . who . is . found . :g\uilty# - of ^murder . by :sh/ooting# . or ^causing an . ex:pl/osion# - ^may be h\anged# - -
(S.5.3.941-951)

In this section, and in others where intonation is commented upon, one misses prosodic marking of the examples. Even the material from the London-Lund Corpus has been stripped of all markers of intonation and most markers of extralinguistic features. Instead, some of the prosody has been reinterpreted and represented by means of punctuation. Sometimes the prosody of invented examples is discussed (e.g. p 169), which I find doubtful. However, on the whole, intonation is shown to be relevant to the use of conjunctionals in speech, particularly in connection with manifestness, which is why it would have been nice to see it included in the exemplification.

Some interesting observations concern the distinction between "normal" causal relations and those in which one state of affairs is the speaker's propositional attitude, as in (5).

5 Has the popstar already gone, because I want to meet her? (invented example, p 115)

The meaning here is "I'm asking you this question because I want to see the popstar". It is found that not all conjunctionals can be used to express this type of attitudinal causal relation. In a comparison of the four registers for the use of formal and attitudinal causality, the Parliamentary oral answers stand out once again by providing over 90% of the total number of attitudinal causal expressions in the whole corpus.

A causal relation can be complex, in that several causes or consequences can be related to the same state of affairs, as in (6). Most, but not all, conjunctionals can be used in this type of construction.

6 Will he also review the whole procedure of the purchase of houses by local authorities so that it may be streamlined and quickened and so that vacant properties may be made available to first-time buyers from local authorities? (POA.15J.446, p 135)

In some cases causal/consecutive conjunctionals seem to have lost most of their causal meaning and function as discourse markers, as in (7).

7 She moved out at the end of April and bought a house with another girl in Acton [...] -- very cheap place. So, you know, well, we we hadn't we'd been scarcely speaking for almost a year, really... (S.2.7.458, p 144)

The conjunctionals that most often assume the function of discourse marker are *because* and *so*. Vandepitte disputes Altenberg's (1984) claim that only these conjunctionals can be used in speech to link larger parts within a discourse, claiming that *for, that's because, consequently, in consequence, that's why, therefore, thus*, and marginally *as* and *since* can have a similar function. Some of these are, however, exemplified only by means of constructed examples (*for, that's because, as, since, in consequence*).

The concluding section of the chapter offers a table (p 149) which summarizes very well the findings presented in the chapter, marking the number of occurrences of the conjunctionals as well as their semantic and pragmatic characteristics.

Chapter IV is concerned with the interpretation of causal relations and with pragmatic acceptability, rather than with the use of conjunctionals. The principle of relevance, with reference to Sperber & Wilson's work, is emphasized as a major factor in the processes of disambiguation and reference assignment. Disambiguation is needed when a conjunctional such as *as* or *since* is used, which can denote a temporal as well as a causal relation. A listener will choose "that lexical meaning specification which involves the least effort [...]. Only if that choice does not yield any contextual effects will it be [abandoned]" (p 158).

The process of reference assignment applies to the identification of the causal relation, as well as to what states of affairs are causally related. For example in (8) the *because*-clause can be related either to "I can only assume", to "she felt", or to "there was some debt of honour".

8 I can only assume that she felt that eh there was some debt of honour, eh, <u>because</u> we had agreed with the Government of China on the terms of the restoration of Chinese sovereignty over Hong Kong. (PI.85, p 161)

The broad view that is taken on conjunctionals and on causal relations is clearly a strong point of Vandepitte's study. It is interesting to see a generative approach to syntax combined with a pragmatic study of register. This multidisciplinary approach enables Vandepitte to treat her topic on a rather full scale. Thus an impressive number of features have been examined which may potentially influence the selection of conjunctional. Some are found to be of importance, while others (such as the distinction between sufficient and necessary cause) are not fruitful.

Although it is admitted that no single parameter works alone to determine the selection of conjunctional, the different influencing factors are kept apart as far as possible in the analysis. There is also a consistent distinction between different levels of language production (linguistic and 'non-linguistic'), which reflects a systematic method of investigation. Importantly, the whole study is conducted in a very open and honest way, so that the reader can follow every step that is taken, and thus be able to witness and evaluate the analysis throughout.

Less impressive, perhaps, is the way the material is handled, as well as the way in which the examples are presented and used. Vandepitte states in the introductory chapter that she does not want to be restricted by the corpus, which is a valid point when one wants to investigate the linguistic system and the borderline between what is and what is not acceptable. Thus, aware that her corpus is limited, Vandepitte often resorts to invented examples in order to illustrate constructions which are not found in her corpus. Although these examples are said to have been checked by native speakers of English, some of them seem distinctly odd, and do not really support the argument. (9) is an example.

9 I knew why I was being vivaed really, probably for I knew I'd done pretty well. (invented example p 49)

The example is there to show that the conjunction *for* can be modified by a modal adverb, which it does not seem able to prove.

The use of invented examples is perhaps particularly doubtful in a study of specific registers. The fact that registers are studied separately at all presupposes that they are different. Thus the fact that a construction is found acceptable in one register does not automatically make it acceptable in another. In this case, the study is concerned with four spoken registers, and the invented examples can hardly be said to be instances of any of those. I also feel that it is especially important that a study of pragmatics should be firmly based on attested language usage. Moreover, even though the invented examples are never included in the tabular surveys of the occurrence and the features of conjunctionals, the authentic and the invented examples seem to have been given equal weight in the process of describing the characteristics of causal expressions.

I find it surprising, in view of the easy availability of computerized corpora with search tools, that Vandepitte does not seem to have consulted other sources for constructions which are infrequent in, or absent from,

her own corpus. On p 84 Vandepitte concludes that certain causal conjunctionals "do not occur -- or only very seldom -- in spoken language", since they are not represented in her corpus. Nevertheless, a quick search in the London-Lund Corpus yielded examples of 4 out of 9 of those conjunctionals, plus one which was related (*accordingly, arising from the fact that, on account of, thereby, with the result that*). Although none of these were frequent, their existence, and the ease with which they were found, illustrate a way in which the need for invented examples could have been greatly reduced.

The use of expressions such as *frequent* and *frequency* is another problem with the treatment of the material, simply because the material is not designed for frequency counts. Each register is represented with the amount of text which was needed to provide 375 examples of causal/consecutive conjunctionals, thus the text samples are not equal in actual size. If the frequency of overtly expressed causal relations in the small-scale study (p 86) is in any way representative, the amount of text in each register varies greatly, with over three times as many words in the Parliamentary oral answers as in the Conversation material. On such a basis one cannot safely claim that a certain construction is more frequent in one register than in another; rather one can predict the likelihood of a conjunctional to be of a certain type whenever causality is expressed.

The reference list of Vandepitte's study is long and comprehensive, and includes literature from many fields. It may thus not seem fair to criticize the absence of any work. However, in spite of the references to Quirk *et al* 1985, which has a similar classification of adverbials, I find it surprising that Greenbaum 1969 has not been consulted; first because this is the work in which the category of conjunct was established, and secondly because it discusses some of the same problems of ambiguous expressions as Vandepitte takes up, such as the conjunctionals *so, hence, therefore, now (that), thus, consequently*; cf Greenbaum 1969: 70 ff.

A clear merit of Vandepitte's study of causal relations is the way in which very different approaches are combined in order to give a broad description of an aspect of language production and interpretation. To me, the treatment of the material, particularly the heavy reliance on invented examples, is disturbing. Nevertheless, this is not crucial for the main argument of the book. Vandepitte has arrived at some interesting conclusions as regards the expression of causal relations by exploring a wide range of syntactic, pragmatic, and cognitive features of such

expressions. She has thus contributed to our understanding of the processes that underlie the expression of causality, and of how linguistic expressions are the result of the interaction of a large number of considerations.

Notes

1. The examples are reproduced here as they appear in Vandepitte 1993.

2. The question mark is used to mark pragmatic unacceptability, in contrast to ungrammaticality.

References

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

Greenbaum, Sidney. 1969. *Studies in English adverbial usage*. London: Longman.

Quirk, Randolph, Sidney Greenbaum, Geoffrey Leech, Jan Svartvik. 1985. *A comprehensive grammar of the English language*. London: Longman.

Studies and Monographs [TiLSM] Book 65) - Kindle edition by Svartvik, Jan. Download it once and read it on your Kindle device, PC, phones or tablets. Use features like bookmarks, note taking and highlighting while reading Directions in Corpus Linguistics: Proceedings of Nobel Symposium 82 Stockholm, 4-8 August 1991 (Trends in Linguistics. Studies and Monographs [TiLSM] Book 65). "Corpus studies and probabilistic grammar". In English Corpus Linguistics ed by K. Aijmer & B. Altenberg, 30-43. London: Longman. Halliday, M.A.K. (1992). "Language as system and language as instance: the corpus as a theoretical construct". Mouton de Gruyter, Directions in Corpus Linguistics: Proceedings of Nobel Symposium 82, Stockholm, 4-8 August 1991, ed. Jan Svartvik (Trends in Linguistics Studies and Monographs 65). Lazarsfeld, P. F. (1962). American Sociological Review, 27(6), 757-767. http:// dx.doi.org/10.2307/2090403. McEnery T., Hardie A. (2012) Corpus linguis...