Mario Brdar
University of Osijek
Faculty of Philosophy

Andrew Wilson, Paul Rayson and Tony McEnery, eds.: *Corpus Linguistics by the Lune. A Festschrift for Geoffrey Leech*. (Łódź Studies in Language 8). Frankfurt am Main – Berlin – Bern – Bruxelles – New York – Oxford — Wien: Peter Lang, 2003. 305 pp.

Before starting the review proper, let me just allow two comments on the title and the subtitle of the volume under review, respectively. The second part of the title of the volume may at first blush appear rather puzzling to some uninitiated readers. One could perhaps get the impression that the syntagma *by the Lune* has ultimately to do with the Medieval Latin *lūna* 'moon', which later appears in French as *lune*. The idea of madness already attaches to the Latin word, later reflected in words such as *lunacy*, *lunatic*, etc. found in various forms in many languages. *Lune* itself is also recorded in OED (2nd edition on CD) as meaning in the plural something like 'fits of frenzy or lunacy, mad freaks or tantrums', which is clearly shown by the following quotations:

1611 Shakes. Wint. T. ii. ii. 30 These dangerous, vnsafe Lunes i' th' King, - beshrew them.

1778 Johnson Let. to Mrs. Thrale 14 Nov., My master is in his old lunes and so am I.

1799 Lamb John Woodvil iii, Let him alone. I have seen him in these lunes before.

Although people doing corpus linguistics, just like computer people on the whole, may be obsessed with computer hardware and software, the former are generally serious but also cheerful and witty people enjoying what they are doing, linguistics included, (cf. Geoffrey Leech as a very good example), so that any obsession with corpora and language phenomena on their part is still within the bounds of what is a healthy scholar attitude.

Or—an equally bad lead—are we to think that the title rather makes some sort of very indirect pun on the development of the root in German, where in Middle High German we find *lűne* developing into the Present-Day German *Laune* 'whim, humour'? But note that on the more positive side, Latin *luna* may be related to roots such as *lux* and *lumen* 'light', as it is claimed that they all come from the same source **leuksna*.

Well, no! These are obviously too far-fetched assumptions, in spite of the fact that corpus linguists may work like mad, and that the results of their analyses may be illuminating. The fact is that the expression *by the Lune* hints at one of the Meccas of corpus linguistics, Lancaster University with its University Centre for Com-

puter Corpus Research on Language (UCREL for short). The name of the town, as recorded in 1086, *Loncastre*, reveals its Roman origins: 'fort on the river Lune.' The Celtic river name may have itself meant something like 'healthy, pure' or be related to Gaelic *Al-non* 'white water'. While the town of Lancaster is rather close to the Lune Estuary, UCREL has always been one of the headsprings of ideas and methods in corpus linguistics ever since Professor Geoffrey Leech founded a group under the name of CAMET (Computer Archive of Modern English Texts) within the Department of English at Lancaster University in 1970 (to become an inter-disciplinary research unit, today's UCREL, in 1984).

This also tells a lot about the subtitle of the volume under review. It is a collection of 15 papers read at the *Corpus Linguistics 2001* conference held at the University of Lancaster (30 March-2 April) to honour Prof. Geoffrey Leech. This is, in fact, his second Festschrift. The first was Thomas and Short (1996), prepared for the occasion of his 60th birthday. Five years later, in 2001, a conference was held to celebrate the life and works of this linguistic giant, who then retired in 2002. Four speakers who have worked closely with Geoff at various stages in his career—Douglas Biber, Jenny Thomas, Geoffrey Sampson and Mick Short—were invited to deliver a special series of lectures during the conference.

Corpus Linguistics 2001 was meant as a forum for all concerned with the computer-assisted empirical analysis of natural language. Among its goals was the encouragement of a dialogue between those working on similar issues in different languages and between areas with a potential to interact, as well as of a dialogue between researchers using corpora in linguistics and those using corpora in language engineering.

The volume under review contains a smaller selection of revised papers from this conference. Another volume of proceedings of the conference, containing 68 full papers and 30 abstracts, was also prepared by the team that edited the present volume (Wilson, Rayson and McEnery 2003).

The papers in the book under review are not ordered thematically but are simply arranged alphabetically according to the surnames of authors. Fortunately, in some cases two, or even three, papers that are related in terms of their topics happen thus to come together.

The first two papers can be considered classical instances of descriptive corpus analyses. The volume opens with a joint paper by Bas Aarts, Evelien Keizer, Mariangela Spinillo, and Sean Wallis entitled "Which or what? A study of interrogative determiners in present-day English." This study of two *wh-* words used as interrogative determiners in noun phrases in independent questions is based on the British component of the International Corpus of English, a one-million-word parsed corpus of 1990 British English, including both naturally occurring spoken and written material that is fully analysed in terms of grammatical tree structures. The data from the corpus seem by and large to square with the statements one can find in most authoritative handbooks on English grammar: NPs with *which* are definite, while those with

what are indefinite. However, Aarts et al. note some instances of divergent usage. It is suggested that the determiner *what* could be considered "a vagueness specifier" (p. 17). While this determiner implies a choice of answers from what appears to be an unlimited set of possibilities. This set may actually have an upper or lower bound, or both. It is claimed that pragmatic factors, such as speaker and hearer expectations, also influence the choice between the two.

This contribution is characteristic of the whole volume in another respect, too. It is, unfortunately, plagued by editorial omissions that may indeed annoy the reader, such as a table broken into two by the page break, or two examples in the middle of the paper, correctly numbered as (13) and (17), respectively, that are referred in the text of the body as 0, or the missing references to two papers mentioned in footnotes (ironically, these are articles written in which two of the three authors were themselves involved).

Karin Aijmer's contribution, "Discourse particles in contrast: The case of *in fact* and *actually*", is concerned with documenting the process of developing meanings and functions of connective particles by the two English modal adverbs of certainty and a range of their Swedish translation counterparts. The former particle is typically used in the corpus for upgrading the evidence for a claim, or when more force is needed in the face of wrong beliefs or expectations. *Actually*, on the other hand, was used in contexts indicating that no such force was involved. Due to their context-dependence and multifunctionality, the two English particles have a large number of translation correspondents. It is claimed that by assuming a contrastive perspective, i.e. by considering their translation possibilities in a parallel corpus we get to know more about the meaning of these particles than we would if we stayed within the bounds of a single linguistic system.

This second paper also shows thematic affinity with the following two papers, which are concerned with more general discourse and historical pragmatic issues in corpus linguistics. Dawn Archer and Jonathan Culpeper ("Sociopragmatic annotation: New directions and possibilities in historical corpus linguistics") show how combining different fields of work, such as pragmatics, historical linguistics, sociolinguistics and corpus linguistics, each with its own methodological assumptions and problems, results in compounded difficulties. Their central problem is how to bridge the gap between text and context. They propose a sophisticated annotation scheme as a remedy to reconstruct the historical social context, and test it on data spanning 120 years (1640-1760) and consisting of 240,000 words from two text-types: trial proceedings and drama. The advantage of their annotation schema is that it allows for the full spectrum of Early Modern English society, incorporating not only classic sociolinguistic variables such as status and age, but also roles as more dynamic aspects of interaction. It also captures the utterance-by-utterance interaction between speakers and their addressees.

The paper by Ylva Berglund and Oliver Mason ("'But this formula doesn't mean anything…!?'") is a report of work on a larger project on the automatic stylistic assessment of L2 essays. The project includes development of suitable tools and

methods for the identification of parameters that reveal the degree of the relative (un-)naturalness of English texts produced by non-native writers.

The two papers that follow focus on lexical phenomena that can be observed in corpora viz. natural language use. First, Doug Biber, Susan Conrad, and Viviana Cortes examine the ubiquity and entrenchment of some pre-fabricated multi-word combinations. As the title of their contribution tells us ("Lexical bundles in speech and writing: An initial taxonomy"), this is one among the first corpus linguistic steps towards documenting most frequent recurring lexical sequences, or lexical bundles for short. These are not necessarily complete structural units or fixed expressions. The paper in this volume concentrates on such bundles involving 3 and 4 words. Their corpus findings directly contradict prior impressionistic expectations: pre-fabricated units are no special characteristic of spoken language. In fact, "[t]he general pattern here is thus one of both similarity and difference…" (p. 82). Both the conversation component and the academic prose component of the Longman Spoken and Written English Corpus (LSWEC) show similar patterns: around 3% of 4 word bundles and 25 % of 3 word bundles in the spoken language, compared with 2% of 4 word bundles and 18% of 3 word bundles in the academic prose (according to the Longman Grammar of Spoken and Written English, p. 933f, as the figures reproduced on p. 76 of the volume under review are unfortunately identical, i.e. the figure of conversation is simply wrong). The two registers, however, differ heavily concerning the structural types of lexical bundles as well as concerning the typical functions of those bundles.

Raymond Hickey's paper entitled "Tracking lexical change in present-day English" is perhaps the least typical corpus study in the volume, but it is nevertheless one of the most exciting ones. It is concerned with the phenomenon of conversion as an instance of univerbation tendencies in English, i.e. of a structural shift by which phrases are reduced to single words. The data is an informal collection of examples attested in American, British and Irish English over a period of two years, either in spontaneous conversation or in the media in the wider sense. Part of the motivation of the stipulated increase in the use of conversions may be attributed to the speakers' wish to achieve greater directness, but also to their wish for increased subjectivity, i.e. for increasing subjective perspective by compacting the expression in such a way that the (1st person) subject and the newly-coined verb come into a contact position.

Barbara Lewandowska-Tomaszczyk, Michael P. Oakes, and Paul Rayson report on the work in progress on the alignment and annotation of a large bilingual corpus intended as assistance with English-Polish translation. It is shown that aligned corpora annotated with part-of-speech and semantic tags may provide invaluable help for translators, enabling them to look up translated portions of text containing specific words, words with certain morphosyntactic or semantic tags.

We not only return to diachronic issues in the papers by Stanley Porter and Matthew O'Donnell ("Theoretical issues for corpus linguistics and the study of ancient languages") and Helena Raumolin-Brunberg ("Temporal aspects of language change: What can we learn from the CEEC?"), but are also invited to think about the

methodological message of these papers. Porter and O'Donnell point out three particular clusters of factors that have to be taken into consideration as steps prior to the compilation of an ancient (and dead) language. These factors have to do with the size and compilation criteria, the annotation procedure and levels of analysis, as well as with the method of analysis. Raumolin-Brunberg uses data from the Corpus of Early English Correspondence (CEEC), compiled at the University of Helsinki, in order to make us aware of how imprecise our reference to time in historical linguistics is and also to point our how many different ways there are for looking at temporal issues. She exemplifies this on five morphosyntactic changes occurring in Late Middle English (c. 1300-1500) and Early Modern English (c. 1500-1700): replacement of subject *ye* by *you*, the replacement of the third-person singular suffix for the present indicative *–th* by *–s*, the transformation of gerunds from an abstract noun to a verbal structure and the concomitant morphosyntactic shift in their argument structure from the *of*-phrase to the direct object, the introduction of the possessive pronoun *its*, and, finally, the loss of the *–n* inflection with first and second-person singular possessive determiners (*mine* and *thine* vs. *my* and *thy*). The difficulties are even more serious if one looks at the temporal aspects of change on the macro- and on micro-levels, i.e. as national aggregates and as individual uses, respectively.

The methodological thread is continued by Geoffrey Sampson ("Reflections of a dendrographer"). He first recounts how he started his work on the database of manually-annotated language samples, specifically, on drawing parse-trees for samples of various genres of language needed to train a statistics-based automatic parser that Geoffrey Leech and Roger Garside had got sponsorship to develop in the 1980s. Such a sample of sentences annotated with labelled trees identifying their grammatical structure was nicknamed "treebank" by Geoffrey Leech. Sampson's team at the University of Sussex has meanwhile extended this work to diverse genres of English, including the spontaneous conversational speech of the CHRISTINE Corpus, and the written output of schoolchildren at Key Stage 2 within the current LUCY project. The over-riding goal of this corpus work is not to produce the largest possible annotated samples of the language, but to achieve transparency and consistency in the primary processing of data in a corpus. Specifically, their treebanks are devised so as to specify the most precise and comprehensive guidelines possible for annotating real-life English, so that annotations scheme should always have a single predictable way of marking the structure. On a more descriptive side, treebanks have made possible some interesting discoveries about English that fall outside anything that could emerge from pre-corpus, intuition-based linguistics, which is exemplified in the paper on a series of mini case studies, including, among others, left- branching structures, tense-aspect system, etc.

Hans-Jörg Schmid asks an extremely intriguing question in his contribution entitled, "Do women and men really live in different cultures? Evidence from the BNC". In trying to answer this question he uses a method borrowed from an earlier work by Leech and Fallon: comparison of frequencies of words and collocations in two demographically different corpora. Actually, in one corpus were all the utterances by males in the spoken section of the British National Corpus, and in the other

all the utterances by females. Findings on words and collocations from domains such clothing, colours, home, food, etc., (where female preponderance was expected), and domains such as swearwords, car, traffic, work, computing, sports, etc. (where male preponderance was expected). As might have been expected, "even perfectly innocuous-looking words are not used with the same frequency by the women and men recorded in the BNC" (p. 210), suggesting that, in keeping with the differences in their prototypical social roles, women and men indeed do not live in exactly the same culture.

The two papers that follow (Mark Sebba and Susan Dray's "Is it Creole, is it English, is it valid? Developing and using a corpus of unstandardised written language", and Mick Short's "A corpus-based approach to speech, thought and writing presentation") are both concerned with the problems of design and compilation of special corpora. The former paper deals with a new written language taking shape in Great Britain—after centuries of being spoken, and only occasionally written down, Jamaican Creole has begun to appear regularly in print in Britain. The paper discusses various difficulties in the course of compiling the Corpus of Written British Creole at Lancaster University. First of all, there is the problem of identifying texts for inclusion. This is in part due to the complex relationship between Jamaican Creole and Standard English, which means that it was not always easy to decide whether a text in question contains Creole linguistic items (and enough of them to be included). A related problem was the range of genres, and the amount of the material belonging to certain genres available for inclusion (informal writing, i.e. unpublished and never intended to be published, was simply not accessible; e.g., personal letters written from one Creole speaker to another). Because only sketchy grammatical descriptions of Creole are as yet available, there were understandable problems in the course of the annotation of the corpus. In spite of all these difficulties and the limited nature of the corpus, its compilation is a significant step to understanding better the nature of written British Creole.

Short's paper is intended as a general report on the progress in a corpus-based investigation of how speech, thought and writing is presented in written and spoken discourse. As corpus-based approach to these phenomena requires precise and systematic analysis, e.g. during annotation, it has helped uncover new phenomena in discourse report and led to useful distinctions not made previously in this sort of research.

Jean Véronis turns in his paper, "Sense tagging: Does it make sense?" to one of the hottest issues in corpora annotation. As the techniques for part-of-speech (POS) tagging have become quite reliable, sense tagging is the next big challenge for corpus linguistics. Véronis points out the discrepancy between the quite substantial body of research on the topic and how little we know about actual human performance in the area. He produces experimental evidence that humans are actually not too good at sense annotation, and that there is a great deal of interannotator disagreement even in relatively simple sense tagging tasks. It is claimed that distributional information coming from the investigation of corpora can provide a sound

foundation for both dictionary organization and sense tagging, without resorting to meaning analysis and the more or less introspective or psychological considerations.

Finally, Anne Wichman and Richard Cauldwell's contribution, "Wh questions and attitude: The effect of context", is an account of methodological problems but also some advantages of approaching the link between intonation and speaker attitude from a corpus linguistic perspective, specifically on the case of *wh*-questions. It is shown that contextual information can change a listener's perception of affect, the most common case being the neutralisation of the perception of negative when context is provided.

In sum, the contributions in this volume clearly show a healthy tendency towards the diversification of corpus linguistic research. On the one hand, we note that new techniques and methodological procedures make it possible to broaden the range of topics that can be approached in this way. While the former sort of broadening of the purview of corpus linguistics may be thought as being primarily on a descriptive level, we also note, on the other hand, a similar, more qualitative diversification whereby some aspects of more or less well known phenomena come to the fore that have as yet gone unnoticed. The wave of these innovations carries with it another, more general broadening of the scope, as new types of specialized corpora are being designed, compiled and tested.

One cannot but agree with G. Sampson when he in the end of his paper sums up his own feelings and, without doubt, the position of the majority of corpus linguists (p. 181f):

> As a corpus linguist at the beginning of the 21st century, I genuinely feel the way that Isaac Newton claimed, surely mock-modestly, to feel:
>
> > like a boy playing on the sea-shore and diverting myself in now and then finding a smoother pebble or a prettier shell than ordinary, whilst the great ocean of truth lay all undiscovered before me. (Spence 1966: §1259)
>
> One thing is sure, though. Even those few pretty shells and pebbles would never have come into my ken if I had stayed inland, working exclusively with sentences like *A ticket was bought by every man*. So it is appropriate in this context to acknowledge the debt which I and many others owe to the man who showed us the path down to the beach. I should like to express my warmest gratitude to Geoffrey Leech.

References:

Biber, Douglas, Stig Johansson, Geoffrey Leech, Susan Conrad, Edward Finegan (1999). *Longman Grammar of Spoken and Written English*. Harlow: Pearson Education.

Thomas, Jenny, Mick Short, eds. (1996). *Using Corpora in Language Research*. London: Longman.

Wilson, Andrew, Paul Rayson, Tony McEnery, eds. (2003). *A Rainbow of Corpora: Corpus Linguistics and the Languages of the World*. (Linguistics Edition 40). München: Lincom-Europa.