REFERENCES

Bowerman, M. (1987). Commentary: mechanisms of language acquisition. In MacWhinney, B. (ed.) *Mechanisms of language acquisition*. Hillsdale. N.J.: Lawrence Erlbaum Associates. 443–466.

Fletcher, P. (1985). *A child's learning of English*. Oxford: Basil Blackwell.

Gleitman, L., Gleitman, H., Landay, B. & Wanner, E. (1988). Where learning begins: initial representations for language learning. In Newmeyer, F. J. (ed.) *Linguistics: The Cambridge Survey*, vol. III. Cambridge: Cambridge University Press.

Hopper, P. J. & Thompson, S. (1980). Transitivity in grammar and discourse. *Lg* **56**. 251–299.

Landay, B. & Gleitman, L. (1985). *Language and experience: evidence from the blind child*. Cambridge, MA: Harvard University Press.

MacWhinney, B. (1987). The competition model. In MacWhinney, B. (ed.) *Mechanisms of language acquisition*. Hillsdale, N.J.: Lawrence Erlbaum Associates. 249–308.

Pinker, S. (1984). *Language learnability and language development*. Cambridge, MA: Harvard University Press.

Radford, A. (1988). Small children's small clauses. *Transactions of the Philological Society* **86**. 1–43.

Zwicky, A. (1970). A double regularity in the acquisition of English verb morphology. *Papers in Linguistics* **3**. 411–418.

Reviewed by PAUL FLETCHER,
Department of Linguistics,
University of Reading.

(Received 18 October 1988)

Douglas Biber, *Variation across speech and writing*. Cambridge: Cambridge University Press, 1988. Pp. xiii + 299.

Douglas Biber's book *Variation across speech and writing*, which is based on his 1984 Ph.D. thesis written at the University of Southern California under the supervision of Edward Finegan, makes important contributions to several areas of linguistic investigation. The areas for which it is most directly relevant include sociolinguistics, computational linguistics, stylistics and text linguistics. Unfortunately, the title of the book does not indicate its broad potential, and it may therefore fail to be noticed by sections of the linguistic community who could benefit from it. To make some amends, I shall in the following briefly outline its significance to the two areas of sociolinguistics and text linguistics. Biber uses a computer to analyze an enormous corpus of spoken and written English language. And it is exactly the computer that constitutes the basis for the progress he is able to make in these fields but also for the work's inevitable limitations.

The database of his analysis is of such enormous proportions that it can only be searched with the help of a computer. Using a tagging algorithm he analyzes large sections of both the London–Lund corpus of spoken English and the LOB corpus of written English and counts the frequency of 67 linguistic features in 481 different texts. He uses only features for which specific discourse functions have been claimed in the relevant linguistic literature. They include, for instance, tense and aspect markers. Past tense is

understood as a surface marker of narrative; the perfect aspect is associated with narrative/descriptive texts; and present tense is seen as a marker of immediate relevance. Other features whose frequency is counted include place and time adverbials, various categories of pronouns, nominalizations, subordinations, various relative clause constructions, downtoners, hedges and discourse particles, type/token ratio, word length, and many more.

However, the limiting factor in the selection of features is again the computer. Place adverbials are a case in point. They are searched for on the basis of a finite list of common adverbs as given by Quirk *et al.* (1985: 516). Biber's list, however, does not include adverbial phrases introduced by *about*, *between*, *in*, *opposite* or *on* because these prepositions 'often mark logical relations in a text' (224). This is of course true, and the omission is systematic across all texts analyzed so that it should not distort the results, but intuitively it seems to leave out a rather large proportion of all place adverbials. In other cases, as for instance the distinction between past tense forms and past participle forms, Biber is forced to post-edit the computer count manually because it cannot distinguish the forms. This means that in spite of the computer processing, there are limits to the overall size of the corpus he can handle.

Biber estimates that his tagging program with the help of post-editing achieves a success rate of at least 90 per cent (217). The unsuccessful items are usually missing tags rather than wrong tags. The first step of the tagging algorithm is to check every word against the vocabulary list of the Brown corpus of American English containing grammatical tags. The actual computer program only comes into force when the Brown corpus contains more than one tag for one item or when it does not contain the item at all. As a consequence of this it leaves untagged 'archaic forms, unusual spellings, or *British spellings*' (220; my emphasis), which may raise an eyebrow or two, considering that the corpora he analyzes contain only British English texts. The actual analysis then clusters the linguistic features into groups of features that co-occur with a high frequency in texts. He establishes seven such groups or factors. These factors are interpreted as textual dimensions on the basis of the shared discourse functions of their individual features.

One example may serve as an illustration. Factor 1 combines 25 features that tend to co-occur with relatively high frequency in certain texts and 9 features that tend to be absent in the same texts but co-occur in other texts. The positive features include *that*-deletion, contractions, present tense verbs, causative subordination, discourse particles, and indefinite pronouns, whereas the negative features include nouns, word length, prepositions, type/token ratio and attributive adjectives (89, 102). The positive features are 'associated in one way or another with an involved, non-informational focus, due to a primarily interactive or affective purpose and/or to highly constrained production circumstances' (105). The negative features, in contrast, are associated with high informational focus. A high type/token

ratio, for instance, indicates a careful lexical choice and a high density of information. Biber (107) summarizes thus: 'Factor 1 represents a dimension marking high informational density and exact informational content versus affective, interactional, and generalized content'. Because more than half of all the features analyzed by Biber are involved in this dimension, he identifies it as a fundamental linguistic dimension.

Factor 2 distinguishes narrative discourse from non-narrative discourse. Factor 3 distinguishes between discourse with highly explicit, context-independent reference and discourse with non-specific, context-dependent reference. Factor 4 is characterized by features that can be labelled as 'overt expression of persuasion'. Factor 5 distinguishes between abstract and non-abstract information, and for factor 6 the label 'on-line informational elaboration' is suggested. Factor 7, finally, is not very strong and is therefore not included in the further analysis, but as an initial label Biber suggests 'academic hedging' (114). The features involved in this last factor include the verbs *seem* and *appear*, downtoners and concessive subordination.

In the next stage of the analysis, Biber computes for each factor a factor score for each text, and for each genre an average score for all its texts. The term 'genre' is taken to refer to text characterizations on the basis of external criteria, that is to say it includes such categories as telephone conversations, press reportages, business letters and so on, whereas the term 'text type' refers to categorizations on the basis of linguistic criteria such as the proposed factors or dimensions. It turns out that each dimension structures the entire set of texts in a different way. On dimension 1, personal telephone conversations are highly involved whereas financial press reportage and natural science academic prose are highly informational. On dimension 2, spot news reportage and political press reportage are narrative whereas technology and engineering academic prose is highly non-narrative. On dimension 3, the technology and engineering academic prose has highly explicit and situation-independent reference whereas personal telephone conversations and sports broadcasts have situation-dependent reference.

This approach then offers a new perspective to text linguistics. Text types are established with rigorously empirical methods on the basis of the frequency of occurrence of linguistic features, something which is only possible with the help of powerful computer programs, which facilitate the analysis of large data bases. They allow the inclusion of a great number of linguistic features, and make it possible to group them into factors according to well-defined mathematical procedures.

A fully fledged typology of English texts clearly has to distinguish between genres and text types while pointing out their interrelationships. Personal letters, spontaneous speeches, and interviews intuitively seem to be rather disparate genres but on the dimension of involved versus informational production, they turn out to be very close together, less involved and more informative than face-to-face conversations but more involved and less

informative than general fiction, press reviews or official documents. On the dimension of narrative versus non-narrative concerns, on the other hand, interviews and face-to-face conversations are very close together, sharing the middle ground between the two extreme poles. It is Biber's contribution to text linguistics to show how a typology of texts on the exclusive basis of linguistic criteria can be achieved. The result turns out to be more complex than might have been expected. There are no clear-cut oppositions of the type [+/−formal] or [+/−involved] or indeed [spoken] versus [written]. What we have to reckon with are several dimensions made up of a group of co-occurring linguistic features. Individual texts or genres are not placed in absolute categories but ranked along the scales of the relevant dimensions.

The computer program yields the factors, that is to say it groups the features according to their co-occurrence patterns, but the interpretation of the factors is still the linguist's job. Here Biber relies heavily on previous research and on the discourse functions that have been claimed for individual linguistic features. For the genres, i.e. the categories assigned to texts on external criteria, Biber relies without qualms on the labels assigned to them by the compilers of the respective corpora. Inevitable as this is in such a research project, it is a problem that deserves careful consideration within a comprehensive typology of texts. Can we ever come up with an exhaustive list of all genres of English, or are they infinite in number? How many genres are recognized by speakers of English? Should the genre classification pay any regard to native speakers' judgments or should it be a matter for the linguist to analyze the non-linguistic features which characterize a particular text or discourse?

The implications of Biber's book for variability studies in sociolinguistics are probably just as far reaching as for text linguistics. What he provides is a meticulous methodology for plotting the variation of a large number of linguistic features across a large number of texts. He does not restrict the linguistic features to items that share 'referential sameness' in order to extract some remaining difference in meaning, either social or geographical, but concentrates on the frequency of linguistic features and takes their discourse functions as the primary elements of his categorization.

He does not consider social or geographical differences within his huge corpus but only genre differences. Given the small range of social classes represented in the two corpora from which he draws his material and the absence of information on the origin of the speakers and writers, this limitation must be accepted as inevitable. But the methodology clearly would be amenable to wider ranging correlational studies of linguistic features with all sorts of extra-linguistic features, provided the appropriate corpora were available.

Biber's book may also, at least to some extent, serve as an introduction to one area of computational linguistics. He is meticulous in introducing the concepts he uses, and in explaining every step in his analysis including the

more advanced techniques in the statistical analysis. On the other hand it is also possible to skip the more technical sections of part II of the book and turn straight to the interpretation of the results in part III, as the reader is indeed encouraged to do (25).

In an appendix, Biber includes brief accounts of the algorithms used to search for every single one of the 67 linguistic features. This leaves him open to attack because a close scrutiny of these algorithms reveals that some short cuts were necessary in order to obtain more or less reliable results, as pointed out above with the example of the place adverbials. But the openness makes the book all the more valuable because it allows a fair assessment of what his figures actually mean, and it indicates where the need for further work is most pressing. The appendix also includes large lists of mean frequencies for all features in each of the 23 genres, and the Pearson correlation coefficients for all the linguistic features. However, if these lists are to be used for comparative purposes, considerable caution is called for, and it seems advisable to check Biber's recognition algorithm very carefully to ensure that the comparison compares like with like.

REFERENCES

Quirk, R. *et al.* (1985). *A comprehensive grammar of the English language*. London: Longman.

Reviewed by ANDREAS H. JUCKER,
Swiss National Science Foundation,
and *University of Zurich*.

(Received 6 January 1989)

Maria Luisa Zubizarreta, *Levels of representation in the lexicon and in the syntax*. Dordrecht: Foris, 1987. Pp. vi + 198.

Zubizarreta's book is set against the background of Chomsky's Government and Binding theory as outlined in Chomsky (1981), and the author relies heavily on modifications to that theory introduced by Williams and van Riemsdijk (1981). I shall not repeat the basic tenets of the theory here, assuming that the reader is familiar with them. The main purpose of my review is rather to offer a general overview of Zubizarreta's own account and to raise a number of general questions.

As the title of the book indicates, the focus of Zubizarreta's interest is levels of representation, and her major modification concerns the role of the lexicon in determining syntactic structure. Essentially, the idea is that not only are there different levels of representation in the syntax, but this proposal must be extended to the lexicon, where she also posits two levels of representation:

(1) (i) S-R: the lexico-semantic level
(ii) L-R: the lexico-syntactic level