# Data, Description, Discourse

## Papers on the English Language in Honour of John McH Sinclair on his sixtieth birthday

Edited by Michael Hoey

HarperCollins Publishers
77-85 Fulham Palace Road
London W6 8JB



First published 1993
**Reprinted 1995**
10 9 8 7 6 5 4 3 2



ISBN 0 00 370947 7

Typeset by Barbers Ltd, Wrotham, Kent
Printed in Great Britain by Bell & Bain Ltd, Glasgow

# Contents

# Acknowledgements

This book would not have been possible without the help of a number of people. I am particularly grateful to Marilyn Washbrook and Kay Baldwin for their help in handling all aspects of the correspondence and many aspects of the manuscript. I am also grateful to Annette Capel, Gillian Hinton, Charlie Ranstead and Richard Thomas for their efficient treatment of this book once it reached HarperCollins. To all the contributors to this volume I extend particular thanks for their promptness and professionalism without which the volume could never have appeared so quickly.

One acknowledgement, however, deserves its own paragraph. Gwyneth Fox planned this volume with me. Together we decided that a festschrift for John Sinclair on his sixtieth birthday was something that we had to undertake and together we decided whom to invite to contribute. The order of the papers and the title of the volume bear Gwyneth's mark as much as my own. In short it was always planned that the volume should come out under our joint editorship. Sadly, however, immense pressure of work coupled with a bereavement meant that in the editorial stages Gwyneth had to withdraw. It is therefore only right to record here that, though the editorial defects of this volume are my own, many of its strengths are properly attributable to Gwyneth.

# Introduction

In a festschrift published in honour of C C Fries many years ago (Marckwardt, 1964), the editor commented in his introduction that Fries had had three ideas in his life; most of us, he added, have only one or two. Despite the favourable contrast with the majority of linguists, the remark struck me then, and strikes me still, as a back-handed compliment and unfair to Fries. Certainly it could never be said of the recipient of this volume. The career of John McH Sinclair, Professor of Modern English Language at the University of Birmingham, has been characterised by an unending stream of original ideas, sometimes carefully worked out in detail, sometimes casually tossed out in papers published in obscure volumes, occasionally developed in large research teams, often passed to a thesis student in the course of conversation. Indeed it is hard to imagine him writing or saying anything derivative or dull, and, while the reader may on occasion be driven to disagree with him, he or she is never tempted to ignore him. It is this ability to think freshly about the English language that the contributors to this volume celebrate, and they do so for the most part in a way that he will appreciate, by offering original ideas of their own.

Not only could it never be said of John Sinclair that he has only had three ideas, it could not even be said that his ideas have all been in the same area of language study. He has made significant contributions in every area of language work except phonology. His work on lexis has been farseeing, both theoretically and practically (e.g. Sinclair, 1966, Sinclair et al, 1970, Sinclair, 1991) and the dictionaries he has edited are recognised to have changed the face of lexicography (e.g. COBUILD 1987, 1988, 1989); he has been involved in writing two quite different kinds of grammar, one of which became a standard native student text and the other of which has become an essential reference work for overseas students (Sinclair, 1972; Sinclair et al, 1990); he has contributed key papers on the nature of text over the years (most recently, Sinclair, 1993); and in the 70's and 80's, in conjunction with colleagues at Birmingham, he established spoken discourse analysis as a major area of linguistic study (Sinclair and Coulthard, 1975; Sinclair and Brazil, 1982).

No-one could accuse John Sinclair of having had only three ideas, but perhaps it would be true to say that he has been driven by three principles. The first of these is that linguistic generalisation should be based on the examination of as much linguistic data as possible. In the sixties when he first articulated this principle it was a strikingly unfashionable thing to say, thanks to the dominance of the generative paradigm with its emphasis on intuition. That it is less unfashionable now is due in part to his repeated demonstration over the years of the obvious value of working with corpora. But working with large corpora brings its own substantial problems, and much of the work of John Sinclair's latter years has

been devoted to developing ways of getting more information out of ever larger computer-held corpora. He has been known to say that the development of the computer with a powerful memory would be to linguistics what the development of a microscope with a powerful lens was to biology - an opportunity not just to extend our knowledge but to transform it. Although it is early days yet, the signs are that he is right.

A second principle has been that a linguist has a responsibility to make clear claims. While the scientific status of linguistics has been much discussed over the years, one aspect that has not always been emphasised is that the so-called 'hard' sciences set up claims in such a way that they can be disconfirmed. This John Sinclair has always tried to do. The hedges, the qualifications, the restrictions, the fuzzy wording are not for him, and in consequence others have been encouraged to use - or argue against - the positions he has laid down. For those who have worked with him, or studied under him, this has been of inestimable importance.

The third principle has been that there is no need to distinguish the descriptive linguist and the applied linguist; there is no conflict of interest between the 'pure' description and the messy application, between the ivory tower and the paddy fields. This is shown in all his major contributions. His work on discourse analysis is a major piece of descriptive linguistics but grew out of a Schools Council project. His recent work on lexicography has resulted in major reference works for language learners but grew out of theoretical and descriptive positions that developed independent of the practical need. In short, John Sinclair has refused to allow himself to be categorised, and the benefits to both descriptive and applied linguistics have been considerable.

The papers in this volume reflect, as one might expect, the range and diversity of interests of the man they celebrate, but they all in different ways reflect the principles just enunciated. M.A.K. Halliday's paper sets the scene by describing a successful attempt to use an eighteen million word corpus of written English, developed by John Sinclair and his colleagues at the University of Birmingham under the auspices of the COBUILD project, to test a long-held belief about grammatical probabilities; the methodology used and the problems encountered are discussed in useful detail. He argues that there is no dichotomy between corpus linguistics and theoretical linguistics and shows that a corpus properly used may support, modify or transform one's theoretical perspective. He also roots current interest in corpus-based description in a tradition going back to the 1950's, an important warning not to allow current excitements to obscure the significance of earlier work.

M.A.K. Halliday's paper is concerned with basic grammatical choices, such as that between past and present tense or between positive and negative polarity, and the research he reports grows naturally out of the systemic tradition with which he is closely associated. Maurice Gross's work, on the other hand, grows out of a quite different tradition - that of transformational generative grammar. It is of interest, therefore, that, despite the difference of perspective, Maurice Gross's paper is also concerned with choice. In particular, he is concerned to model a feature of the language that causes problems to grammarians and lexicologists alike, namely that of idiomatic expressions. The question he engages with is: how can one represent formally (i.e. on a computer) the structure of such expressions so as to reflect their relatedness and at the same time their differences. Making use of the notion of a

'local grammar', he shows how the 'synonymy' of certain expressions can be handled with a combination of finite state graphs and transformations. This way of handling such expressions is likely to have substantial effects on the shape and scope of current parsers.

In his concluding remarks, Maurice Gross notes that lexically frozen structures are more numerous than free ones. The cline from free structures through collocations to idioms is one that anyone who works with a corpus is bound to attend to. Stig Johansson's paper, like Maurice Gross's, is interested in the relationship of lexis and grammar, focusing in particular on the variety of meanings that an adverbial-adjective combination may have. Using data drawn from the tagged LOB corpus, he demonstrates that while the grammatical description of these combinations may be straightforward their semantic patterning is much more complex. Relating the different uses of adverbial modification that he finds exemplified in the data to current descriptions, he shows how corpus data can extend and alter our understanding of apparently well-understood phenomena.

Like Stig Johansson, Gerhard Leitner is interested in the relationship between grammatical description and semantic patterning; he explores this relationship, however, from a different perspective. Building on John Sinclair's and M.A.K. Halliday's view that the only difference between grammar and lexis is one of degree of delicacy, he examines the grammatical treatment of the lexical items *start* and *begin* in three recent dictionaries. He shows how the dictionaries in question differently handle the grammar of these common words and notes defects (and strengths) in each treatment. His conclusion that there is a need for a semantically-based approach to English grammar and for dictionaries and grammars that mesh emphasises the close relationship that has to exist between good theory and description and good practice, a relationship as noted above that John Sinclair has always strived to maintain.

All the papers so far mentioned are concerned to use a computer-held corpus or computational modelling in order to say something about the English language as a whole. It is however possible to employ similar techniques of lexical and grammatical investigation in order to say something about the nature of the corpus itself. In other words, the corpus becomes the object of investigation rather than the tool. The next three papers in this collection all seek to answer two questions explicitly articulated by Michael Stubbs and Andrea Gerbig at the beginning of their paper: 'What patterns of meaning exist across long texts, and what methods are available for describing and interpreting them?' and 'How can an individual text be located in diatypic space relative to other texts, text types and text corpora?'

Michael Stubbs and Andrea Gerbig examine a school geography text-book from a variety of points of view, including lexical density, the grammatical encoding of human activities and the representation of human agency, and the techniques they demonstrate are described in sufficient detail to permit others to engage in similar analysis. Despite some similarities with the techniques used in the research reported by M.A.K. Halliday in this volume, the objective of Michael Stubbs and Andrea Gerbig in their paper is rather different (though not one which M.A.K. Halliday would find uncongenial): the ways in which texts may construct social reality. They conclude that the grammatical encoding of the book they examine hides the labour of working people and de-emphasises causation. They comment that textual

analysis makes ideological structures tangible and that the development of appropriate computational methods (to which they themselves have contributed) offers the possibility of uncovering unsuspected patterns of language which in turn will reveal something of the construction of social reality.

At several points in their paper, Michael Stubbs and Andrea Gerbig compare the frequency of certain features in their chosen school textbook corpus and the large, general LOB corpus. They also note that interpretation does not spring mechanically from statistics. Malcolm Coulthard's paper likewise compares the frequency of certain features found in his specialist corpus with those found in a general corpus and expresses similar caution about the relationship between the features one discovers in a corpus and any interpretation one might make of those features. The difference is that Malcolm Coulthard's data are linguistic objects of dispute in the courtroom, and interpretation of findings with regard to them may have an immediate legal consequence such as contributing to a conviction or acquittal. Malcolm Coulthard notes the need for comparing disputed texts not only with large general corpora but with smaller specialist corpora as closely like the texts under investigation as possible. Thus the 'norm' is defined in two ways - as the general language with which the specialist language contrasts and as the specialist language with which the individual text may contrast. He also notes the need for more research into determining 'how small a sample one can reliably work with, at what size the significant regularities begin to emerge' and so on. This is in keeping with John Sinclair's third principle; a practical application (forensic linguistics) leads to important theoretical and descriptive questions.

Angele Tadros's paper has features in common with both Malcolm Coulthard's and Michael Stubbs and Andrea Gerbig's papers, in that, like the latter, she seeks to relate the textual features observed to the view of the world projected by the writers and in that, like the former, she seeks to compare small and closely similar specialist corpora. She concludes that a writer's decision to cite or not depends on the relationship s/he seeks to establish with his or her readers. But the ability of the reader to recognise what the writer is assuming about that relationship is in part dependent upon the reader having reached a certain level of sensitivity to the voices in a text. Angele Tadros notes that a student needs to be able to discriminate between the author's voice and the voices of those the author invites into the text s/he is creating.

One of the pieces of text that Angele Tadros analyses in her paper is a lengthy extract from one of John Sinclair's most important works, co-authored with Malcolm Coulthard, *Towards an Analysis of Discourse* (1975). My own paper treats that book not as data but as the starting point for arguing that an extra rank is needed in discourse above the exchange. I argue that some analytical problems disappear if the unit of 'exchange complex' is posited.

In my paper I hint that the resources of the exchange complex may be differently utilised according to the 'genre' of interaction. In his paper Ronald Carter notes that recognising conversational styles should be part of the 'knowledge about language' that a pupil acquires at school. He argues that pupils should be made more aware of the way in which language varies according to its function. This is part of a much larger discussion of the importance of developing different kinds of language awareness amongst pupils. He suggests how pupils might be encouraged to apply what they find out about language to newspaper and literary texts and makes

a strong case for learning about language being a necessary part of the curriculum, commenting that 'learners are better learners if they are able to analyse what they are doing and *why* they are doing it'.

As director of the LINC (Language in the National Curriculum) project, Ronald Carter is aware of the importance of linguists maintaining a satisfactory dialogue with government, although his paper does not deal with that relationship directly. Robert Kaplan's paper is however partly about aspects of the relationship between government and language experts. Reporting on an important language planning consultancy in New Zealand, he notes that New Zealand poses some complex planning problems. He describes in some detail the New Zealand language situation with its legal recognition of Maori as the official language and its de facto acceptance of English as the official language, and provides a full account of the range and kinds of language used in the country. He concludes that if New Zealand is to preserve its multilingual diversity it will have to act fast, since 'language death ... is not a pretty thing to watch'.

At one point in his paper, Robert Kaplan comments: 'It is difficult to be very specific about the language situation in New Zealand because no linguistic data bases have ever been maintained', and throughout his paper, he takes care to show where more information is needed and on what basis his conclusions are drawn. This indeed is the common thread that runs through all the papers in this volume. They are all concerned with describing some facet of the English language carefully on the basis of adequate data, whether that facet be lexical, textual, discoursal, educational or sociolinguistic. In their variety and in the qualities they share, the papers of this volume are appropriate tributes to a man whose own work has been varied but who has always respected data, offered tight descriptions, and placed his data-driven descriptions within a discoursal context.

## References


*Collins COBUILD English Language Dictionary* (1987) London: HarperCollins Publishers.

*Collins COBUILD Essential English Dictionary* (1988) London: HarperCollins Publishers.

*Collins COBUILD Dictionary of Phrasal Verbs* (1989) London: HarperCollins Publishers.

*Collins COBUILD English Grammar* (1990) London: HarperCollins Publishers.

Marckwardt, A.H. (ed) (1964) *Studies in Language and Linguistics in Honor of Charles C. Fries*, Ann Arbor: English Language Institute, University of Michigan.

Sinclair, John McH. (1966) 'Beginning the study of lexis' in Bazell, C.E. et al (eds) *In Memory of J.R.Firth*, London: Longman, pp 410-430.

Sinclair, John McH. (1972) *A Course in Spoken English: Grammar*, Oxford: Oxford University Press.

Sinclair, John McH. (1991) *Corpus, Concordance, Collocation*, Oxford: Oxford University Press.

Sinclair, John McH. (1993) 'Written discourse structure' in Sinclair, J.M. et al (eds) *Techniques of Description: Spoken and Written Discourse* pp.6-31.

Sinclair, John McH. & Brazil, David (1982) *Teacher Talk*, Oxford: Oxford University Press.

Sinclair, John McH. & Coulthard, R.M. (1975) *Towards an Analysis of Discourse*, Oxford: Oxford University Press.

Sinclair, John McH.; Jones, S. & Daley, R. (1970) *English Lexical Studies*, OSTI Report, University of Birmingham.

# Selected Publications of John Sinclair

1965 'When is a Poem like a Sunset?', *Review of English Literature* 6, 76-91, and in Lyle E B (ed.) *Ballad Studies*, Brewer/Rowman and Littlefield, reprint, 1976.

1966 'Taking a Poem to Pieces' in *Essays on Style and Language* ed. Roger Fowler, London: Routledge, 68-81, and in *Linguistics and Literary Style*, ed. Donald M Freeman, New York: Freeman, Holt, Rinehart and Winston.

1966 'Beginning the Study of Lexis' in Bazell, C.E.; Catford, J.C.; Halliday, M.A.K, and Robins, R.H. (eds) *In Memory of J.R. Firth* London: Longman, 410-30.

1968 'English Language in English Studies' *Educ. Rev.* 20, 82.

1968 'Linguistics and the Teaching of English' *Language and Language Learning* ed. A H Marckwardt, NCTE, Champaign, Illinois, 31.

1968 'A Technique of Stylistic Description', *Language and Style* 1, 215-242.

1970 (with S Jones and R Daley) *English Lexical Studies*, Birmingham University, for the Office for Scientific and Technical Information.

1971 'The Integration of Language and Literature in the English Curriculum', *Educ. Rev.* 23, 220-234.

1972 *A Course in Spoken English: Grammar*, Oxford: Oxford University Press

1972 (with R M Coulthard, I J Forsyth and M C Ashby) 'Discourse in the Classroom', London: Centre for Information on Language Teaching Research.

1972 (with M C Ashby, R M Coulthard and I J Forsyth) *The English used by Teachers and Pupils*, Final Report to the Social Science Research Council for the period September 1970 to August 1972, Birmingham: University of Birmingham.

1972 'Lines about Lines', in Kachru B; Stahlke B; Herbert F W (eds) *Current Trends on Stylistics*, Edmonton, Linguistic Research Inc. 251-61, and in Carter R (ed.) *Language and Literature*, London: Allen & Unwin 163-76, 1982.

1973 'English for Effect' *Commonwealth Educ. Liaison Com. Newsletter* 3, no 11. 5-7. reprinted in Stubbs, M. and Hillier, H. (eds) *Readings on Language, Schools and Classrooms*, London; Methuen, 1983, pp 238-245

1973 'Linguistics in Colleges of Education', *Dudley Journal of Education* No 3, 17-25.

1974 'English Lexical Collocations' *Cahiers de Lexicologie*, Paris: Institut des Professeurs de Français a l'Etranger.

1975 'The Linguistic Basis of Style', Ringbom, H. et al (ed) *Style and Text*, Stockholm: Språkforlaget Skriptov AB and Abo Akademi in collaboration, 75-89

1975 (with R.M. Coulthard) *Towards an Analysis of Discourse: the English Used by Teachers and Pupils*, Oxford: Oxford University Press.

1979 'Issues in current ESP Design and Management' in S Ziahosseiny and A Mountford (eds) *English for Special Purposes*, papers from the 2nd Regional ESP Conference, Isfahan.

1980 'Discourse in Relation to Language Structure and Semiotics'; in Greenbaum S; Leech G and Svartvik J (eds), *Studies in English Linguistics for Randolph Quirk*, London: Longman, 110-24.

1980 'Computational Text Analysis at the University of Birmingham' in Johansson, S (ed.) *Newsletter of the International Computer Archive of Modern English*, The Norwegian Computing Centre for the Humanities, Bergen.

1980 'Applied Discourse Analysis: an Introduction', in Sinclair, J M (ed) *Applied Linguistics*, 1, No 3, Clarendon Press, Oxford, 253-61.

1980 'Language for Specific Purposes' in Kennedy C (ed.) *English Language Research Journal* No 1, University Of Birmingham, 3-13.

1981 'The Development of Skills for Learning' in Blackie D (ed) *English for Specific Purposes* 47, 1-5 Oregon State University.

1982 'The Integration of Language and Literature in the Curriculum' in R Carter and D Burton, (eds) *Literary Text and Language Study*, London: Arnold.

1982 (with D. Brazil) *Teacher Talk*, Oxford: Oxford University Press, the first part of J. Sinclair reprinted as *The Structure of Teacher Talk*, Discourse Analysis Monographs No 15, Birmingham: English Language Research, University of Birmingham.

1982 'Reflections on Computer Corpora in English Language Research', in *Computer Corpora in English Language Research* (ed) Stig Johansson, Bergen.

1982 'Linguistics and the Teacher', in Carter R (ed) *Linguistics and the Teacher*, London: Routledge and Kegan Paul, 16-30.

1983 'Planes of Discourse', in S N A Rizvil (ed) *The Two-fold Voice: Essays in honour of Ramesh Mohan*, Pitambur Publishing Co., India, 70-91.

1984 'Lexicography as an Academic Subject', in Hartmann, R K K (ed.) *LEXeter 83 Proceedings* Lexicographic Series Maior No 2, Tübingen: Max Niemeyer Verlag, 3-12.

1984 'The Teaching of Oral Communication', *Nagoya Gakuin Daigaku Gaikokugo Kyoikukiyo*, 10, 1-12.

1984 'Naturalness in Language', in Aarts J and Meijs E (ed) *Corpus Linguistics: Recent Developments in the use of Computer Corpora in English Language Research*, Rodopi, Holland, 203-10; and in *Ilha do Desterro: A Bilingual Journal of Language and Literature*, Vol. V, No 11 45-55 Florianopolis, Brazil, 1985; and in *ELR Journal* No. 2, University of Birmingham.

1984 'Language Awareness in Six Easy Lessons', in B G Donmall (ed) *Language Awareness*, NCLE Papers and Reports 6, National Congress on Languages in Education, 4th Assembly.

1984 (with Donmall, B G and Tinkel A J) 'Evaluation and Assessment' (in *NCLE Papers and Reports* 6, National Congress on Languages in Education, 4th Assembly).

1985 'On the Integration of Linguistic Description' in van Dijk T A (ed) *Handbook of Discourse Analysis*, vol. 2, 13-28. London: Academic Press.

1985 'Lexicographic Evidence' in Ilson R (ed) *Dictionaries, Lexicography and Language Learning*, ELT Documents 120, Pergamon.

1985 'Basic Computer Processing of Long Texts' in Candlin C and Leech G (eds) *Computers and the English Language*, London: Longman.

1986 'Fictional Worlds' in Coulthard R M (ed) *Talking about Text*, Discourse Analysis Monograph No 13, ELR, University of Birmingham.

1986 'First Throw Away Your Evidence' in Leitner G. (ed) *The English Reference Grammar*, Tübingen: Max Niemeyer Verlag.

1987 (ed) *Looking Up*, London: HarperCollins Publishers.

1987 'Grammar in the Dictionary' in Sinclair, J M (ed) *Looking Up*, London: HarperCollins Publishers.

1987 'The Nature of the Evidence' in (ibid).

1987 'The Dictionary of the Future', Collins English Dictionary Annual Lecture, University of Strathclyde. Reprinted in the *Library Review* and *Focus on English*, British Council, Madras, 1987.

1987 'Language Models and Monuments' in *Britain Abroad*, The British Council Magazine, No 3.

1987 'Tools of the Trade' in *TESOL France Newsletter*.

1987 (with A Renouf) 'A Lexical Syllabus for Language Learning' in Carter R and McCarthy M (eds) *Vocabulary and Language Teaching*, London: Longman.

1987 'Collocation: A Progress Report' in Steele R & Threadgold T (eds) *Language Topics: Essays in honour of Michael Halliday*, Amsterdam: John Benjamins.

1987 (editor in chief) *Collins COBUILD English Language Dictionary*, London: HarperCollins Publishers.

1987 'Compressed English' in M Ghadessy (ed) *Varieties of Written English*, London: Pinter.

1988 'Mirror for a Text' in *Journal of English and Foreign Languages* No 1 Hyderabad, India.

1988 'Sense and Structure in Lexis' in Benson J; Cummings M; Greaves W (eds) *Linguistics in a Systematic Perspective*, Amsterdam: John Benjamins.

1989 'Language, Learning and Community' in Candlin, C and McNamara (eds) *NCELTR* Macquarie University (originally published in 'Corpus Creation' paper presented to Council of Europe, February 1987).

1990 (with D Kirby) 'Progress in Computational Lexicography' in *World Englishes* Vol. 9, No 1 Oxford: Pergamon.

1990 'The Nature of Lexical Statement' in *Linguistic Fiesta,* a festschrift for Professor Kakehi, Japan: Kurioso Publishers.

1990 'Uncommonly Common Words', in Tickoo M (ed) *Learners' Dictionaries: State of the Art,* Anthology Series 23, RELC Singapore.

1990 'Methods and Madness' in Bickley V. (ed) *Language Use, Language Teaching and the Curriculum,* Hong Kong: Institute of Language in Education.

1990 'Teaching English: the Decade Ahead' in Abousenna M (ed) *Proceedings of the Tenth National Symposium on English Language Teaching,* Cairo: CDELT.

1990 (editor in chief) *Collins COBUILD English Grammar,* London: HarperCollins Publishers.

1991 *Corpus Concordance Collocation,* Oxford: Oxford University Press.

1991 "The Automatic Analysis of Corpora" in *Proceedings of the Nobel Symposium on Corpus Linguistics*, Stockholm.

1991 (with A Renouf) 'Collocational Frameworks', in Aijmer, K and Altenberg, B (eds) *English Corpus Linguistics: Studies in Honour of Jan Svartvik,* London: Longman.

1992 'Trust the Text' in Davies M & Ravelli L (eds) *Advances in Systemic Linguistics: Recent Theory and Practice,* London: Pinter.

1992 'Shared Knowledge' in *Proceedings of the Georgetown Round Table.*

1992 'Priorities in Discourse Analysis' in Coulthard, (ed) *Advances in Spoken Discourse Analysis,* London: Routledge.

1993 'Written Discourse Structure' in Sinclair, J M; Hoey, M, and Fox, G (eds) *Techniques of Description: Spoken and Written Discourse,* A Festschrift for Malcolm Coulthard, London: Routledge, 6–31

1993 (with M Hoey and G Fox) (eds) *Techniques of Description: Spoken and Written Discourse,* a Festschrift for Malcolm Coulthard, London: Routledge.

# Biographical Notes on Contributors

M.A.K. Halliday is Emeritus Professor of Linguistics at the University of Sydney. His major work has been in systemic functional grammar and its use in language description, text analysis and language education. Among his recent publications are *An Introduction to Functional Grammar* (Edward Arnold, 1985), *Spoken and Written Language* (Oxford, 1989), and [with J.R. Martin] *Writing Science: Literacy and Discursive Power* (Falmer, 1993).

Maurice Gross is Professor of Computational Linguistics at the University Paris 7. He is also the Head of the Laboratoire d'Automatique Documentaire et Linguistique of the French National Center for Scientific Research. He is the author of *A Transformational Grammar of French* (4 volumes).

Stig Johansson is Professor of English Language at the University of Oslo. He is coordinating Secretary of the International Computer Archive of Modern English (ICAME) and editor of the *ICAME Journal*.

Gerhard Leitner is Professor of English Language at the Free University of Berlin. He has published and edited widely on reference grammars of English: *The English Reference Grammar,* Tübingen: Niemeyer (1986); *Reference Grammars and Modern Linguistic Theory,* Tübingen: Niemeyer (1989); *English Traditional Grammars. An International Perspective,* Amsterdam: Benjamins (1991). He has also published on varieties of English in Britain, Australia and India.

Michael Stubbs is Professor of English at the University of Trier, Germany. He previously worked at the University of Birmingham (1973-74), Nottingham (1974-85), and London (1985-90) where he was Professor of English at the Institute of Education. He was Chair of BAAL (1988-91). He has published articles and books on educational linguistics and on text and discourse analysis.

Andrea Gerbig is a Research Associate at the University of Trier, Germany. She graduated from the University of Duisburg, and has worked on research projects in Australia and Namibia. She is currently writing a Ph.D. on the computer-assisted analysis of a corpus of texts on the environment.

Malcolm Coulthard is a Senior Lecturer in English Language at the University of Birmingham. His interests include spoken and written discourse analysis and forensic linguistics. His recent publications include *Introduction to Discourse Analysis*, 2nd edition (1985), the edited volume *Advances in Spoken Discourse Analysis* (1992) and (in Portuguese) *Linguagem e Sexo* and *Traduçao: Téoria e Práctica* both published in 1991.

Angele Tadros, Associate Professor, was Director of the English Language Servicing Unit (ELSU), University of Khartoum, Sudan. She is now in the Department of English, College of Arts, King Saud University, where she teaches English and Applied Linguistics. Her interests include discourse analysis, the analysis of students' interlanguage and the teaching of writing. Her recent publications include 'Predictive Categories in University Text Books' in *English for Specific Purposes*, 8, 1, 1989.

Michael Hoey is a Senior Lecturer in English Language at the University of Birmingham. His interests include spoken and written discourse analysis and lexis. He has published several books, most recently *Patterns of Lexis in Text*, and many articles on various aspects of discourse.

Ronald Carter is Professor of Modern English Language in the Department of English Studies at the University of Nottingham. He has published widely in the fields of Applied Linguistics, Literary Stylistics and English Language Education. From April 1989 – March 1992 he was National Coordinator of the Language in the National Curriculum (LINC) Project. He is currently editor of the Routledge *Interface* series on Language and Literature and a new applied linguistics series for Penguin Books with David Nunan. He is co-editor with Professor John Sinclair for the Oxford University Press *Describing English Language* series.

Robert B Kaplan is Professor of Applied Linguistics and past Director of the American Language Institute at the University of Southern California. He is the Editor-in-Chief of the *Annual Review of Applied Linguistics* which he founded in 1980; he is also a member of the Editorial Board of the *Oxford International Encyclopaedia of Linguistics* (responsible with H G Widdowson for all Applied Linguistics entries). He has published 25 books and approximately 100 articles. He is currently President-elect of the American Association of Applied Linguistics (AAAL).

# QUANTITATIVE STUDIES AND PROBABILITIES IN GRAMMAR

*M.A.K. Halliday*
University of Sydney

## Part I

At a recent conference devoted to modern developments in corpus studies, I was struck by the way that a number of speakers at the conference were setting up an opposition between "corpus linguists" and "theoretical linguists" – not a conflict, I mean, but a distinction, as if these were members of two distinct species. I commented on this at the time, saying that I found it strange because corpus linguistics seemed to me to be, potentially at least, a highly theoretical pursuit. Work based on corpus studies has already begun to modify our thinking about lexis, about patterns in the vocabulary of languages; and it is now beginning to impact on our ideas about grammar. In my view, this impact is likely to be entirely beneficial. Corpus linguistics brings a powerful new resource into our theoretical investigations of language.

One consequence of the development of the modern corpus is that we can now for the first time undertake serious **quantitative** work in the field of grammar. Quantitative studies require very large populations to work with. This is obvious in lexis, where the underlying principles of the frequency distribution of lexical items have been known for quite some time (since the ground-breaking studies by Zipf (1935), in fact): if we want to investigate any words other than those of highest frequency in a language, we need text data running at least into millions of words, and preferably into hundreds of millions. Even to study the most frequent words, once we start investigating their collocational patterns, we need very large samples of text. It might be argued that grammatical patterns do not demand text data on that scale, because they will typically occur more often. That is true of the broadest, primary categories, like singular and plural in the English noun, or positive and negative in the verb. But the finer, more delicate our categories become, the less frequently each instance will occur; and even with the broader categories, since many of the ones we are likely to be most concerned with are categories of the clause and above, it will require a rather large sample (if we are thinking in terms of the number of words) to yield a sufficiently large number of occurrences. Consider for example the **clause nexus**, a structure of two clauses related by expansion or projection: Nesbitt & Plum were able to retrieve 2,733 instances from a corpus of 100,000 words. If we want to compare the grammar of different registers, the functional varieties of a language, in quantitative terms (for example, the different proportions of active and passive in different registers of Modern English), then it is clearly going to require a very large corpus of data to produce reliable results.

I myself first became interested in grammatical frequencies in a crude and simplistic fashion many years ago in the course of my work as a language teacher – although, if I pursue this even further, the reason for my interest in such information as a teacher was that I had felt the need for it in the first place as a learner. Faced with the task of learning from scratch, as a young adult, a language very distant from my own, or from any language I had had experience of before, namely Chinese, I was constantly wanting to know what people actually said in this language. When I had to master very basic systems of mood, aspect, phase and so on that were unlike anything in English, I wanted to know which of a set of alternatives was the most likely – I did not know the term "unmarked" at that time, but what I was looking for was some guidance about which, if any, was the unmarked term. A few years later, when I started to teach that same language, although I had by then acquired some feeling for its patterns of discourse, I found it frustrating that I could not offer the students reliable information of the kind that I myself would have liked to have.

In other words, I wanted to know the **probabilities** that were associated with these grammatical choices. Given, for example, a particular grammatical system, say that of **aspect**, with terms **perfective/imperfective**, what was the relative probability of choosing one or the other in a Chinese clause? This is a considerable problem for a speaker of a western European language, for two reasons. On the one hand, the aspect system itself is unfamiliar: there is some form of grammatical aspect in English, the exact nature of which is disputed, but the dominant temporal system in the clause is tense, whereas in Chinese the dominant temporal system is that of aspect. On the other hand, there is the problem of the nature of the unmarked term. Grammatical systems in Chinese typically embody an alternative which is "unmarked", not in the sense of being the default or most frequent choice (it often will be that too), but in the sense that it is a way of opting out – of not choosing either of the terms. So the aspect system is not simply "either perfective or imperfective" but "perfective, imperfective **or neither**". Speakers of European languages are much less accustomed to systems of this kind; and again, it would be helpful to know the relative probability of choosing one way or another.

Of course, every learner will carry over into the new language some predictions based on the experience of their mother tongue, and maybe also of other languages that they know. Some of these predictions will hold good: I would imagine that the ratio of positive to negative clauses is substantially the same in all languages, although it would be nice to know. But some of them will not hold good; and there will be some cases where the learner has no idea what to predict at all. And this is often where one begins to reflect on these matters; as long as the predictions work, they tend to remain unconscious. But there was another question which kept arising in my work as a teacher, especially in preparing for conversation classes; and this was something I found quite impossible to predict: was the probability of a choice in one system affected by a choice in another? Could I combine freely, say, negative polarity with perfective aspect; or different voice-like options (the Chinese *bǎ* construction, for example) with interrogative as well as with declarative mood?

When I wrote my first sketch of a grammar of Chinese (1956), I attached probabilities to most of the primary systems. These were, obviously, in a very crude

form: I used the values 0+ ½-, ½, ½+ and 1-. The values were derived mainly from my own knowledge of the language, backed up from two sources: a small amount of counting of grammatical options in modern Chinese dramatic texts; and the work that I had subsequently been doing on a text in early Mandarin, the fourteenth century Chinese version of the *Secret History of the Mongols*, in which I had counted every instance of those grammatical categories that I had been able to resolve into systems. One reason for doing all this counting had been to try to establish the extent of the association between different systems: to find out, for example, whether it was possible to predict the number of instances of negative interrogative by intersecting the probabilities of negative (versus positive) with those of interrogative (vs. declarative). On the basis of this quantitative work, although obviously I was able to access only very minute samples from the modern language, I adopted the principle that frequency in text instantiated probability in the system.

Any concern with grammatical probabilities makes sense only in the context of a paradigmatic model of grammar, one that incorporates the category of **system** in its technical sense as defined by Firth[1]. The system, in Firth's system-structure theory, is the category which models paradigmatic relations. Just as a **structure**, in Firth's specialized use of the term, is a deep syntagm, so to speak, so a system is a deep paradigm. The system, as Firth put it, "gives value to the elements of structure": it specifies the oppositions, or sets of alternatives, to which a defined place in structure provides the condition of entry. Firth's own application of these concepts was largely confined to phonology; but, if we want to give a brief illustration from grammar, using modern terms, we could say that the element "Finite operator" in the structure of the English verb (verbal group) is given value by the systems of **verbal deixis** (temporal/modal) and **polarity** (positive/negative) which originate there. This concept of the system enables us to show that, under given conditions, the speaker is selecting one, and only one, from among a small closed set of possibilities; and therefore it makes sense to talk about the **probability** of choosing one or the other, What I hoped to do was to model each system not just as "choose *a* or *b* or *c*", but as "choose *a* or *b* or *c* **with a certain probability attached to each**". In other words, I was positing that an inherent property of any linguistic system is the relative probability of its terms.

I have often encountered considerable resistance to this idea. People have become quite accustomed to lexical probabilities, and find no difficulty in accepting that they are going to use *go* more often than *grow*, and *grow* more often than *glow* (or, if you prefer a semantically related set of words, that they will say *go* more often than *walk* and *walk* more often than *stroll*). They do not feel that this constitutes any threat to their individual freedom. But when faced with the very similar observation that they are going to use active more often than passive, or positive more often than negative, many people object very strongly , and protest that they have a perfect right to choose to do otherwise if they wish. And of course they have; that is exactly the point that is being made. They could choose to use negative more often than positive, just as they could choose to use *stroll* more often than *walk* – but they won't. The resistance seems to arise because grammar is buried more deeply below the level of our conscious awareness and control; hence it is more threatening to be told that your **grammatical** choices are governed by overall patterns of probability.

Before being able to pursue these studies any further with Chinese, however, I changed the focus of my work and found myself professionally involved with English. So again I started counting things, this time using an early version of a system-based grammar of the language first worked out in collaboration with Jeffrey Ellis and Denis Berg, and subsequently elaborated together with John Sinclair, Angus McIntosh and others at the University of Edinburgh in the early nineteen sixties. I collected a small sample of four different registers of English, just big enough to yield a total of 2,000 occurrences of whatever category provided the entry condition to the systems I wanted to study. For example, in order to count instances of indicative / imperative mood, I had to have 2,000 independent clauses, because it is here that the choice is made: each independent clause selects one or the other. But to compare declarative with interrogative I had to count 2,000 indicative clauses, because it is the indicative clause that is either declarative or interrogative. The reason for settling on a figure of 2,000 occurrences was the following: first, I estimated it needed about 200 occurrences of the less frequent term to ensure a reasonable degree of accuracy; and second, that the less frequent term in a binary system seemed to occur about 10% of the time. So if I restricted the counting to binary systems, 2,000 instances tended to yield around 200 occurrences of the less frequent term in the system.

The systems that I was interested in were mainly clause systems, although this did not imply that every clause would select in the system in question: each system would have its origin in some specific class of the clause (but a very large class, like the "indicative" referred to above). So from each of four registers (a science magazine, a novel, a conversation and a play) I took samples that would be large enough to yield 500 instances of whatever category was required. The systems I investigated were nine in all. The first eight were: (1) voice (active/passive), (2) transitivity (transitive/intransitive), (3) tense (simple/complex), (4) theme (unmarked/marked), (5,6) mood (indicative/imperative, and, within indicative, declarative/interrogative), (7) polarity (positive/negative), (8) nominal deixis (specific/non-specific). The ninth was the system of tone, one of the systems of intonation in English; here there is a separate unit as point of origin of the system, namely the tone group, and five terms (falling/rising/ level (low rising) / fall-rise / rise-fall). For this I used a recorded conversation containing about 1,500 tone groups.

Such samples were of course extremely small, much too small for serious quantitative work, and in any case I was far from confident in the grammatical categories I was using as the basis, because these had not yet been put to the test; indeed one of my main reasons for doing this kind of close analysis of the data was to test the validity of the categories themselves when applied to natural texts, including spoken texts. At this time a very sharp distinction was being drawn in linguistics between the system and the instance (the text), or between competence and performance; and quantitative effects were dismissed as "merely performance features", so very few people were interested in this sort of study (and there was certainly no question of anyone publishing the results!). However it seemed to me that these were important issues, very relevant to our overall understanding of language itself, and that some interesting patterns seemed to emerge. One which I reported on at the time was this. There seemed to be certain systems where the frequency of the terms was more or less equal: given a binary system (and as already noted I had confined the counting to systems that could be represented as binary,

apart from the system of tone), this meant a ratio of about fifty : fifty, or (in terms of probabilities) of 0.5 : 0.5. There were then other systems where the frequency was unequal. But these latter were not distributed across the whole range of possible values. They seemed to tend – again, very roughly – towards a skewing by about one order of magnitude, a "ten to one" ratio. This I represented as 0.9 : 0.1. The former type were those where, from the point of view of frequency, there was no unmarked term; the latter were those in which one of the terms was unmarked.

I did not attach much weight to this observation, for obvious reasons: the whole procedure was far too unreliable. But I did wonder about whether such a pattern would make sense. We can imagine two possible alternatives: one, that the probability profiles of different systems might be randomly distributed across all possible values, from equiprobable to highly skew; the other where the skew systems might cluster around a particular value, but not at 0.9 : 0.1 – say at 99 to 1, or 3 to 1. It seemed easy to suggest that, in some vague sense, 99 to 1 would be too much skewed to be useful in the grammar of a language, while 3 to 1 would not be clearly enough distinguishable from even probabilities; but such an explanation would hardly stand up by itself. At the time, I was just beginning to investigate child language development, as a result of working with primary school teachers on an educational programme concerned with initial literacy; and since I had first faced up to the issue as a language learner, it seemed natural to try to look at it in developmental terms. I had not yet begun to make any direct observations of my own on how children learn their mother tongue; but later when I came to do this, I was struck by the way they respond to the more frequent options in the grammar and use these first, bringing in the less frequent ones by a later step. In other words, children seem to learn language as a probabilistic system; they are surrounded by large quantities of data, probably at least a hundred thousand clauses a year, and they are sensitive to relative frequency as a resource for ordering what they learn to say. (I am not suggesting they do any of this consciously, of course!) From this point of view, one could hypothesize that a semiotic in which the probabilities associated with various sets of options, or systems, were distributed randomly all the way from 0.5 : 0.5 to 0.99 : 0.01 would be virtually impossible to learn. One in which there was some kind of bimodal distribution, on the other hand, would be much more accessible to a learner. This did not in itself favour one particular profile over another for systems of the type which were skew, but it did suggest that that they might very well tend to cluster around just one set of values.

Among the systems I had counted at the beginning was tense. What I was counting here was not the opposition of past and present, or past and non-past; it was that of "simple tense" versus "complex tense". This was based on an analysis of the English tense system different from that favoured by the structuralist linguists, which had only past and present tense (rejecting the traditional notion of future) and combined these with "continuous" (or "progressive") and "perfect" as forms of aspect. My own analysis was more in harmony with the traditional concept of tense, and as interpreted in semantic terms by Reichenbach. In this view, tense (past / present / future) is a potentially iterative system, construing a "serial time", in which there is a primary tense choice of past, present or future relative to the moment of speaking, and also the option of then making a further choice, where the time frame of the primary tense is taken as the reference point for another tense selection of past, present or future – a secondary time system that

is relative to the primary one. So as well as simple past *took*, simple present *takes* and simple future *will take*, there is the possibility of **past** in past, *had taken*, **present** in past, *was taking*, and **future** in past, *was going to take*; likewise **past** in present *has taken*, **present** in present *is taking*, **future** in present *is going to take*; and **past** in future *will have taken*, **present** in future *will be taking*, **future** in future *will be going to take*. This second time choice can then serve as reference point for a third, and so on. In this analysis, a verb form such as that in *I hadn't been going to tell you* is "future in past in past"; that in *he was going to have been looking after things for us all this time* is "present in past in future in past". In all, sequences of up to five tense choices have been observed to occur[2].

Both this and the structuralist form of analysis can be used to throw light on the English tense system, and each will illuminate a different facet of it. The reasons for using the iterative form of analysis would take too long to go into here, particularly as they require illustration from discourse – this analysis treats tense more as a discourse feature. The relevance to the present discussion is that, when interpreted in this way, tense becomes an instance of a very general kind of system found in language: the kind of system that embodies an iterative option, a choice of "going round again". Another example of this would be projection (direct and indirect speech and thought), where we can say not only *Henry said that the day would be fine* but also *Mary reported that Henry had said that the day would be fine, Peter forgot that Mary had reported that Henry had said that the day would be fine*, and so on. Such systems obviously require very large samples for counting the relative frequency of sequences of different lengths. The only one I tried to count was tense, which is less demanding because it is a system of the verb (verbal group) and so each tense form, however complex, remains within the limits of one clause. Comparing simple tense forms (primary tense only) with complex tense forms (primary plus secondary), I found that there was the same general pattern of skew distribution: simple tenses were about ten times as frequent as complex ones. In other words, having made one choice of tense, you can go round and choose again; but you are much more likely not to – and roughly in the same proportion as you are more likely to choose positive than negative or active than passive. This kind of iterative system is something that would appear highly prominent to a child learning the language, and it seems to provide a kind of prototype or model of a skew system, as being a system having one option that is highly marked (much less frequent than the other).

So while my original interest in the quantitative aspect of grammatical systems had been an educational one – its importance for learning and teaching languages – in the course of working on texts, first in Chinese and then in English, I had become aware of the different information loading that different systems can carry. Now, in the late nineteen fifties I had had the privilege of working alongside Margaret Braithwaite, together with A.F. Parker-Rhodes and R.H. Richens, in the pioneering Cambridge Language Research Unit, one of the early projects concerned with machine translation. In this context it became necessary to represent grammatical features in explicit, computable terms. I wanted to formalize paradigmatic relations, those of the system; but I did not know how to do it – and I totally failed to persuade anyone else of this! The emphasis at that time – and for many years to come – was on formalizing syntagmatic relations, those of structure, which seemed to me less central to the task. One of the lines of approach that I

tried to explore was through Shannon and Weaver's (1949) information theory. This had already been rejected by linguists as being of no interest, partly because Shannon and Weaver's own incursions into language had been purely at the orthographic level (what they meant by the "redundancy of English" was the redundancy of the system consisting of twenty-six letters and a space), but partly also because of the current obsession with structure. Information theory has nothing to say about constituent structure; information and redundancy are properties of systems. But it does provide a very valuable measure of the information content of any one system.

A binary system whose terms are equiprobable (0.5 : 0.5) has an information value of 1 (H = 1); redundancy is 1 - H, so it has a redundancy of 1 - 1, which is zero. Redundancy is therefore a measure of the skewness of a system; the greater the skewness (departure from equiprobability), the lower the value of H (information) and consequently the higher the redundancy. (This is, of course, "information" and "redundancy" in these specific mathematical senses, without any implication that one or other is to be avoided!) The original Shannon and Weaver formula for measuring information was $-\sum p_i \cdot \log_n p_i$ where n is the number of terms in the system (so $\log_2$ if the system is binary) and $p_i$ is the probability of each. I used a table of the values of $p_i \log_2 p_i$ for p = 0.01 to p = 0.99, in order to calculate the information of grammatical systems with different degrees of skewness. But when I reported on this it aroused no kind of interest, and when I came to count the frequencies of the systems referred to above, in texts in English, I did not have enough confidence in the figures (or in my own interpretation of them) to pursue the question of information and redundancy any further.

Meanwhile in the 1960s the foundations were being laid for an eventual breakthrough in our understanding of linguistic systems, both qualitatively and quantitatively: namely the beginning of corpus-based linguistics. The "corpus" had been the brainchild of two leading specialists in English linguistics: Randolph Quirk, in Britain, and Freeman Twaddell in the United States. By the middle of the decade the "Brown Corpus" at Brown University in Providence and the Survey of English Usage at University College London were making available bodies of text data that were sufficiently large to allow valid quantitative studies to be carried out. A groundbreaking piece of work was Jan Svartvik's *On Voice in the English Verb* (1966), in which Svartvik used the text resources of the Survey of English Usage to investigate the use and significance of the passive in written English. Since that time a steady output of research has come from these corpus projects, the high point of this achievement being the "Quirk grammar". At the same time other corpus studies were being undertaken; for example in lexis, by John Sinclair and his colleagues in Birmingham, and in grammar by Rodney Huddleston and the members of the "scientific English" project in my own department at University College London. Work of this kind clearly demonstrated the value of this general approach to the study of language[3]. By the 1970s the corpus was well established as a research resource, and was being extended into the domain of language education, both as a resource for foreign language learning (for example the "Leuven Drama Corpus" compiled by Leopold Engels at the Catholic University of Leuven), and as a theoretical basis for work in initial literacy (e.g. the "Mount Gravatt Corpus" developed by Norman Hart and R.H. Walker at Mount Gravatt College of Advanced Education in Brisbane).

As far as quantitative studies were concerned, the corpus entirely transformed the scene. On the one hand, samples of text were becoming available that were large enough for reliable statistical investigation into features both of vocabulary and of grammar. On the other hand, these texts were now in machine-readable form, so that step by step, as appropriate software was developed, it was becoming possible to handle such large bodies of data in a way that would have been impossible with any form of manual processing. Meanwhile one critical contribution to these studies came from the sophisticated methodology for quantitative analysis developed by William Labov, together with Gillian Sankoff and David Sankoff, in their investigation of socially conditioned variation in the phonology and morphology of urban speech (see for example, D. Sankoff, 1978). This has been adapted to systemic-functional corpus studies in syntactic and semantic variation, where (unlike Labov's work) what is being investigated is systematic variation in patterns of meaning, on the plane of content rather than the plane of expression.

These studies are well known, and I have discussed them in a similar context elsewhere (Halliday, 1991, 1992). What follows is a very brief summary. Plum & Cowling (1987) used the corpus of interviews from Barbara Horvath's *Sydney Social Dialect Survey* to study variation in the system of temporal and modal deixis in the English verbal group, and within temporal deixis the system of past / present / future primary tense. They examined 4,436 finite clauses, and found that 75% selected temporal deixis and 25% modal; while of those selecting temporal deixis, leaving out a very small proportion of futures, they found 57% selecting past and 43% selecting present. These were from texts in one particular register: spoken interviews in which the respondents were asked to recall their childhood experiences in primary school, especially the games they used to play. Examining the data for systematic variation within the population, Plum & Cowling found no significant variation in the choice of tense versus modality; but, within tense, they found significant variation among three social groups: relatively, the middle class favoured past tense in their narratives (70% : 30%), the lower working class favoured present tense (65% : 35%), while the upper working class fell in between, but closer to the middle class figures (60% of clauses past).

Using a different corpus, spoken interviews with dog fanciers discussing the breeding and showing of dogs, Nesbitt & Plum (1988) examined a sample of 2,733 clause nexuses to investigate the internal relationship between two systems within the grammar: interdependency (parataxis / hypotaxis) and logical-semantic relations (expansion / projection, and their sub-types). Here the question concerned the intersection (mutual conditioning) of the two sets of probabilities: were the two systems independent, or were they associated in some way? It would take too long to summarize their findings; but let me mention one of them. In the intersection of interdependency with projection (the combination of "parataxis / hypotaxis" with "locution / idea" which defines the four categories traditionally referred to as "direct and indirect speech and thought"), they found that there was a strong association of parataxis with locution ("direct speech") and of hypotaxis with idea ("indirect thought"); both "indirect speech" and "direct thought" were highly marked combinations. In other words, there was a strong conditioning of the probabilities within the grammar itself; and it remained constant whichever system was taken as the environment of the other.

On a considerably larger scale, since the mid-1980s Ruqaiya Hasan has been conducting research into conversation between mothers and children, where the children were just below school age (3½ - 4 years). Her aim has been to investigate how children's patterns of learning are developed through ordinary everyday interaction in the home, and to explore the consequences of this early semantic experience for their subsequent learning in school. Using a population of 24 mother-child dyads, structured by social class (12 "higher autonomy professions", 12 "lower autonomy professions") and sex of child (12 boys, 12 girls), Hasan and her colleagues collected over 60,000 clauses of natural conversation, of which they analysed over one third in terms of a detailed network of semantic features. Subjecting the results to a cluster analysis brought out some highly significant correlations between the social factors of class and sex on the one hand and the orientation towards certain patterns of semantic choice on the other. For example, while all mothers used a great deal of reasoning, middle class and working class mothers tended to favour different kinds of grounding for their explanations, while mothers of boys differed from mothers of girls in the ways in which they elaborated on their answers to their children's questions[4].

Let me take up one final point before moving to the second part of the paper. In 1980 William Mann, director of the "Penman" artificial intelligence project at the University of Southern California Information Sciences Institute, asked me to write a grammar for use in text generation by computer. I constructed a network, on systemic principles, for the grammar of the English clause, based on the work I had been doing since the mid 1960s; there were 81 systems in the network. Whenever possible I represented these as binary systems and added probabilities to them, using just the two values of 0.5 : 0.5 and 0.9 : 0.1 that I had arrived at earlier. This network was then implemented computationally by their programmer Mark James. The network was of course designed to be used under the control of some higher level system, a "knowledge representation system" of some kind; but for testing it was operated randomly. When let loose in this way it produced garbage, as such grammars always will until sufficiently constrained. But when it was operated still randomly but with the probabilities taken into account, Mann's comment was that, while it still produced garbage, the garbage now looked as if it might bear some distant resemblance to English. That may not sound to you very encouraging, but that remark did more than anything else to persuade me to reactivate my interest in probabilities[5].

I thought again about this bimodal effect, of probabilities tending towards either 0.5 : 0.5 or 0.9 : 0.1, and formulated it as a hypothesis about the typology of grammatical systems: that they fall into one or other of these two types, the "equi" and the "skew", with the "skew" having a value of very roughly nine to one, or one order of magnitude in our decimal scheme of things. Then, wondering again about this nine to one, I looked once more into the Shannon & Weaver formula for calculating information. We saw that, at 0.5: 0.5, H = 1, R = 0. What, I wondered, was the point at which information and redundancy exactly balance out (H = 0.5, R = 0.5)? – the property of fifty percent redundancy that Shannon & Weaver had originally assigned to "English" (meaning by that the English writing system, based on the relative frequencies of the twenty-six letters of the alphabet and the space). It turns out that, in a binary system, the probability profile where information and redundancy match one another, at fifty percent each, is almost exactly 0.9 : 0.1. To

give the exact probabilities to two places of decimals: at probabilities of 0.89 : 0.11, H (information) = 0.4999. In other words, the particular distribution of probabilities to which these skew systems seemed to conform was that where there is fifty percent redundancy in the system. This could perhaps begin to suggest an explanation for this kind of phenomenon in the grammar – if it turned out to survive under adequate large-scale scrutiny. To investigate it properly, it was necessary to have access to a sufficiently large corpus of modern English.

## Part II

During the 1980s John Sinclair launched the COBUILD Project in Lexical Computing at the University of Birmingham in England. COBUILD stands for "Collins Birmingham University International Language Database". The task of this project was to assemble a large corpus of modern English text and to develop a range of software that would make this text available for lexicographical purposes – specifically, for research into English lexis that would lead to the production of a dictionary, or a series of dictionaries. At the same time, the corpus would provide a data base for various other purposes such as the production of reference grammars and teaching materials.

Sinclair and his colleagues compiled a corpus of twenty million words of contemporary English: British English, largely written, but including two million words of speech (mainly speech of median level of formality such as interviews on BBC radio). The selection of texts was determined by various factors; but one guiding principle was that it should be in some sense "typical", the sort of English that a learner would want to understand. In this way any material based on the corpus would be useful for students of English as a second or foreign language[6].

The output of the project in its first phase is now well known. It included a series of English teaching materials and a reference grammar; but above all it included the COBUILD series of dictionaries, recognized around the world for their highly innovative approach to defining, classifying and contextualizing English words. The fundamental feature of all these various publications is that they are corpus-based: not only in that all cited instances are taken from the corpus, but in that the overall profile that is presented of each word is that which is revealed by studying how the word is actually used. The project has now entered its second phase, in which the dictionaries will be revised and updated and other publications such as a dictionary of collocations will be put in hand. For this purpose, the original corpus of 20 million words has now been supplanted by one of two hundred million, covering a wider and more structured repertory of registers and dialects of English and including a more substantial component of texts in spoken language. The 200 million word corpus is known as the "Bank of English".

In 1991 I was able to spend four months at the University of Birmingham, working with the COBUILD project; and in particular, during that time I was fortunate in being able to collaborate with one of their research staff, Zoe James, on problems of quantitative grammar: the quantitative analysis of grammatical systems through investigation of frequencies in the corpus. Zoe and I decided to work with the original COBUILD corpus; the larger, new corpus was not yet fully accessible, and in any case 20 million words seemed enough for what we were trying to achieve. In the event we decided to use just the 18 million words of written text, so

that later on, if we wanted, we could compare the results with an equivalent quantity of spoken. So the findings that I shall be presenting in the rest of this paper are derived from a corpus of eighteen million words of modern written English. The results are set out in Halliday and James (1993).

The first question that we had to decide, obviously, was what to count. What grammatical systems should we use as our field of investigation? Here there were a number of factors that had to be kept in view.

1. The systems to be counted had to be very general ones, not those of a more "delicate" kind; they should be systems that apply to a large number of instances. This is partly to ensure that each term occurs with sufficient frequency; but there is a more significant factor, which is that any general hypothesis about probabilities ceases to apply when one moves to more specific sets of options, because these tend to have complex entry conditions. A system network is not a strict taxonomy. For example, we can make a prediction about "polarity: positive/negative" because this system is a very general one, entered by every member of the class of major clauses. But if we were interested in the system of "negative transfer" (the contrast between a pair of expressions such as *I think it's not fair* and *I don't think it's fair,* or *you're required not to smoke* and *you're not allowed to smoke*, or *you mustn't smoke here* and *you can't smoke here*), not only would the population be hard to find but there is no prediction to be made about the relative frequency, because only clauses which are **both** negative **and** modalized enter into this system, and there is no way of knowing how far these two factors condition one another.

2. The features to be counted must, obviously, **be systemic**: that is to say, they must be such that, given some population *x*, we can ask how many have the feature *a* and how many have the feature *b*. That sounds very simple. But suppose there exists another population *y* that has feature *a* by default? Are these to be counted as instances of *a* or not? This is actually a very common phenomenon: for example, in English, a transitive verb (verbal group) can select either active or passive; but an intransitive one is active by default – so are intransitives to be included in the category of "active", or not? A countable noun (nominal group) is either singular or plural; but a mass one, which is uncountable, is singular by default. A strict version of the hypothesis would **reject** instances of *y,a* (intransitive active, or mass singular) because in system-structure theory (as formulated by Firth) what we are calling feature "*a*" cannot, in fact, **be** the same feature under those two different circumstances – it must have a different value in a context where it is in contrast with feature *b*, as opposed to a context where it is not. Eventually, of course, we would want to count such patterns both ways; but there is no point in doing this until the hypothesis has been tested in the simpler cases, those where this problem does not arise.

3. The systems should be ones that are highly loaded semantically, that do a large amount of work in the grammar. In fact this follows from 1. above, because systems which are very general in application will *ipso facto* be highly loaded. But this also suggests that they should be systems of higher rank (systems of the clause, rather than systems of the group or the word), because clause systems occupy a larger semantic domain. For example, positive / negative characterize a whole process, whereas singular / plural characterize only one participant in a process.

4. The systems should be such that we could formulate and test the hypothesis already outlined: that is to say, to start with there should be one system which was predicted to be "equi" and one which was predicted to be "skew". And both should be binary systems, because for a system with any greater number of terms the notion of "skew" becomes more complex: if there were three terms, these might be conceived as 0.9 : 0.05 : 0.05, or as 0.45 : 0.45 : 0.1, as well as other more convoluted profiles. With a binary system there was only one interpretation for "skew".
5. And finally, the system must be recognizable: that is, it must be such that instances of each term could be clearly identified in the corpus. This requirement overshadows all the rest.

Anyone who has tried to write a grammar, or a part of one, in computable form (and that means more or less any grammarian, these days, because thanks to Chomsky we have a much stronger sense of commitment to making our grammars explicit) is aware of the triviality trap: that the easier a thing is to recognize, the less it is worth while recognizing it. Any system that plays an important part in the grammar is likely to be more or less "cryptotypic", construed through a complex string of realizations and/or field of reactances. This is not simply Murphy's law at work; on the contrary, it is precisely because fundamental contrasts like that between, say, material, mental and relational processes have such complex realizations and reactances that they can exert such a powerful role in the construction of meaning. But whereas in my small-scale counting in the 1960s I had been able to count by inspecting every instance myself ("manually", as they say!), now we had to be able to get the programme to do the counting. Each instance had to be recognizable by the computer ("mechanically"). The text was entered in the machine in ordinary orthography, exactly as printed; and it was indexed by the orthographic form of each word. There were no clause boundaries or other grammatical boundary markers, and no grammatical analysis had been carried out on the text. So the systems had to be such that we could formulate some strategy for recognizing every instance in the corpus and assigning it to one feature or the other, with a reasonable chance of getting it right. We could only ever approximate to the true picture; but we wanted to be able to approximate to it with an acceptably low margin of error.

With these considerations in mind, we decided to tackle two systems, as follows:

1 Polarity: positive / negative

2 Primary (deictic) tense: past / present (see (2) below)

The hypothesis was that polarity would be skew, with 0.9 positive : 0.1 negative; while primary tense would be equi, with each term around 0.5.

What was the relevant population for these systems, as represented in a systemic grammar of English? In principle, all **major** clauses choose for polarity, whether finite or non-finite, whereas only **finite** clauses have primary tense, and not even all of those, because some finite clauses select modal and not temporal deixis. Furthermore, of those that select temporal deixis there are some which select neither past nor present but future. So our two systems did not have the same conditions of entry.

This does not matter. There is no reason why the two systems being investigated should have identical populations. But certain decisions had to be taken.

(1) We decided to eliminate non-finite clauses from the polarity count as well as from the primary tense count. This was because non-finites are hard to identify: for every occurrence of an *-ing* form of a verb it would have to be determined whether it was functioning as Predicator in the clause or not.

(2) We decided to eliminate "future" from the count for primary tense. This was because we wanted to treat tense also as a binary system. (In fact instances of primary future were also included in the count; we discuss the possible relevance of this at the end of Halliday and James (ibid.))

(3) Imperatives are anomalous: they are non-finite verbal groups (and hence have no primary tense), but realize finite clauses (as is shown by the possibility of adding a mood tag, as in *Keep quiet, will you!*). This means that they would belong in the polarity count but not in the primary tense count. We decided, in fact, to treat them where they belonged.

So our populations were:

**polarity:** all finite major clauses **including** imperatives, futures and modals;

**primary tense:** all finite major clauses **excluding** imperatives, futures and modals.

[It turned out that the exclusion of these three categories reduced the population by about 16%.]

What we needed to do, then, could be summed up briefly as follows:

**A. Polarity:**

1. identify and count all finite clauses;
2. within this set, identify and count all those that are negative;
3. subtract the negative from the total and label the remainder "positive";
4. calculate the percentage of positive and negative within the total set of finite clauses.

**B. Primary Tense:**

1. (as A) identify and count all finite clauses;
2. within this set, identify all those that
   (i) are non-indicative (= imperative),
   (ii) have modal deixis (i.e. Finite operator is a modal), or
   (iii) have future tense deixis (i.e. Finite operator is temporal: future), and eliminate from the set;
3. within this reduced set, identify and count all those that are past;
4. subtract the past from the total and label the remainder "present";
5. calculate the percentage of past and present within the reduced set of finite clauses.

This could be considered as an idealized research strategy for the investigation. But that, of course, was the point: it was idealized and quite remote from anything resembling reality. It depends on operations like "identify a clause", "identify a

finite clause", "identify a negative clause" and so on; and these things cannot be done automatically. There are of course parsers in existence, including two or three systemic parsers in different stages of development, but even if any of them turned out to be **accurate** enough, they are nowhere near to being **fast** enough. After all, we are talking about a population of 18 million words, which we would expect to yield between one and two million finite clauses.

What is needed is a system for recognizing patterns: some kind of pattern specifier, or pattern matcher, which is specifically designed for the purpose, and which works on the principles of the existing battery of software. So we set out to design this together, with me saying what I thought needed to be done and Zoe working out how to do it – or else saying it couldn't be done and we'd have to think of something else. Now there were two possible approaches here. One would have been to try and work out all the steps in advance and them implement them all at once, counting the results at the end. The other was to proceed step by step, by progressive approximation, counting up as we went along. We decided on the latter course, so that we could monitor the problems encountered and the results obtained at each successive step.

As a first step, we began by listing all the finite verbal operators in English. These form a closed class, and one which is reasonably small: 43 members, when we restricted them to single word forms (that is, including *can't, couldn't* and *cannot*, but excluding *can not* and *could not* – all such forms are regarded by the program as two separate words and would therefore have to be retrieved in a separate operation). We took the decision to exclude the words *need* and *dare*, although these do occur as Finite operators; *dare* is fairly rare in the corpus overall, while *need*, though common, occurs mainly as a "full" verb (that is, it is construed lexically not grammatically). The 43 Finite operators are shown in Table 1 (taken from Halliday & James, 1993). They are set out in two columns by 21 rows; but the rows fall into three blocks, so the Table is structured as two columns by three blocks. The **columns** show polarity: positive on the left, negative on the right. The **blocks** show the operators organized by categories of verbal deixis: at the top, modal, and future temporal; in the middle, present temporal; and at the bottom, past temporal. Those in the top block are those eliminated from the count of primary tense.

Table 1 also shows the results of this initial counting. The total number of instances of positive polarity come out at somewhere under a million, those of negative at between 60 and 70 thousand. Let me make one or two observations regarding this Table.

(1) Some occurrences of the **positive** forms will have been instances of a lexical verb, not an operator: namely **some** occurrences of all forms of *be, have* and *do* (for example, *does* in *he does his homework after tea*, of which the negative is *doesn't do*, not just *doesn't* as it would have been if *does* had been an operator). We shall see below that this does not in fact matter, **provided that such forms do not get counted twice.**

(2) The wordings *is to / was to, has to / had to* (e.g. in *who is to know?, he had to be told*), express modality; so it might seem as if they should not be counted as present and past temporals. But their **deixis** is, in fact, temporal not modal; that is to say, **as Finite operators** they do express primary tense, so they are in their appropriate place.

Table 1. Finite Operators

| POSITIVE | | | | NEGATIVE | | |
|---|---|---|---|---|---|---|
| | NUMBER OF INSTANCES | SPURIOUS INSTANCES OUT OF 200 | CORRECTED COUNT | | | |
| can | 33,002 | 0 | 33,002 | can't, cannot | 9,568 | |
| could | 32,164 | | 32,164 | couldn't | 4,102 | |
| may | 17,716 | 7 | 17,096 | mayn't | 3 | |
| might | 12,509 | 1 | 12,446 | mightn't | 85 | |
| will | 34,817 | 10 | 33,076 | won't | 2,818 | modal, and |
| would | 48,882 | | 48,882 | wouldn't | 3,094 | future |
| shall | 3,787 | | 3,787 | shan't | 137 | temporal |
| should | 16, 053 | | 16,053 | shouldn't | 700 | |
| must | 15,520 | 0 | 15,520 | mustn't | 296 | |
| ought | 1,547 | | 1,547 | oughtn't | 40 | |
| | | | 213,573 | | 20,843 | |
| am | 6,499 | 3 | 6,401 | ain't | 476 | |
| is | 149,514 | | 149,514 | isn't | 3,037 | |
| are | 70,925 | | 70,925 | aren't | 1,160 | |
| have | 76,207 | | 76,207 | haven't | 1,533 | present temporal |
| has | 33,749 | | 33,749 | hasn't | 501 | |
| do | 32,011 | 0 | 32,011 | don't | 16,737 | |
| does | 7,387 | | 7,387 | doesn't | 3,069 | TOTAL PRESENT: |
| | | | 376,194 | | 26,513 | 402,707 |
| was | 186,839 | | 186,839 | wasn't | 4,144 | |
| were | 60,276 | | 60,276 | weren't | 839 | past temporal |
| had | 109,835 | 8 | 105,442 | hadn't | 2,415 | |
| did | 19,322 | | 19,322 | didn't | 9,637 | TOTAL PAST: |
| | | | 371,879 | | 17,035 | 388,914 |
| | | TOTAL POSITIVE: | 961,646 | TOTAL NEGATIVE: | 64,391 | |

**Results:** For Polarity, 961,646 clauses were found to be positive and 64,391 negative out of a total of 1,026,037. For Primary Tense, 402,707 clauses were found to have present primary tense and 388,914 were found to have past primary tense out of a total of 791,621.

**Note:** The "corrected count" is arrived at by deducting from the "number of instances" a number extrapolated from the figure of disqualified occurrences in a sample of 200. For example, out of the 200 instances of *may*, 7 were found to be other than Finite operator (the month of May, in some context or other). The raw count of 17,716 was therefore reduced by 7/200 (i.e. 3.5%) to 17,096. If no disqualified occurrences were found, the figure is shown as 0. Where there is no entry in the column, the test was not applied.

(from M.A.K. Halliday & Z.L. James, 'A quantitative study of polarity and primary tense in the English finite clause', in John M. Sinclair, Michael Hoey and Gwyneth Fox (eds.), *Techniques of Description: Spoken and Written Discourse*, London: Routledge, 1993).

(3) The system cannot distinguish, in counting occurrences of a word, between pairs of homographs, for example *can* as modal and *can* meaning 'sealed metal container'. Since this problem was going to crop up all the time, we considered

various possible solutions. One was to use a tagger; there are two taggers available in the COBUILD repertory, as well as others that have been developed elsewhere, e.g. by Geoffrey Leech at the University of Lancaster. But we reasoned that, even with the fastest tagger, tagging 33,000 instances of *can* is a lengthy business; and we were not convinced of the level of accuracy that could be attained. We adopted an alternative strategy: that of inspecting a random sample of 200 instances, one by one, counting the number that were "spurious" (not instances of the form we were looking for) and then extrapolating from this figure to arrive at a proportion of the total which was to be discarded. The results of this procedure are shown in the column headed "spurious instances out of 200". This is obviously a source of inaccuracy; there were no tin cans in the 200 instances of *can*, for example, whereas no doubt there would have been some in the total of 33,000 or so. On the other hand, 1.5% of occurrences of *am* were found to be 'a.m.' (morning), written without full stops; and this was probably slightly above the average figure. But we reckoned that the margin of error was tolerably small.

If these figures were then taken as a first approximation to the picture for polarity and primary tense, what were the main errors and omissions that were involved? We identified four major sources of inaccuracy, as follows.

1. Mood tags (e.g. *haven't you?*, in *You've seen it, haven't you?*) have been counted as separate instances. The fact that they do not form separate **clauses** does not matter – we are not counting clauses; so since they (typically) retain the tense of the clause, including them will not distort the proportions of past and present. But they do not retain the polarity – in fact they typically reverse it. The polarity system in mood tags is not an opposition of "positive/negative"; it is an opposition of "polarity reversed / polarity constant", with "polarity reversed" as the unmarked option. Hence, since the majority of clauses are positive, counting the tags separately will have **over**estimated the figure for **negative**.

2. All abbreviated Finite operators have been omitted. There are six of these: *'s 'd 've 're 'm 'll*. This may have affected the figures for primary tense, since only *'d* is past, whereas *'s 've 're* and *'m* are all present. It has certainly distorted the figures for polarity, since all those omitted are positive. This therefore has **under**estimated the figure for **positive**.

3. Clauses with negative **words**, such as *not, no, never, nothing,* have been counted (if at all) as positive (for example *the idea was no good, nothing would have helped*). But at least some of these are negative clauses, as shown up by the tag: *the idea was no good, was it?* This has **both over**estimated the **positive** and **under**estimated the **negative** count.

4. But undoubtedly the major source of inaccuracy is the following: that not all English finite verbs have (distinct) Finite operators. Simple past and simple present tense, in positive declarative active form, have their finiteness fused with the lexical verb, as in *I think so, I thought so*; and these are likely to be the two most frequently occurring forms of the verb. They have been omitted altogether; they are all positive (except as under no. 3 above); so this has **under**estimated the total figure for **positive**.

We tried therefore to take steps to deal with each of these errors and omissions. The way we did this was as follows.

(1) We identified the mood tags by counting all occurrences of the sequence comma + operator + personal pronoun + question mark e.g. , *won't they?* It turned out, when we checked this later, that it should have been "comma + operator + personal pronoun + any punctuation mark", since a tag may be followed, in written English, by full stop, exclamation mark or comma, as well as by a question mark. But the total number of tags was quite small (almost certainly under 3,000), no doubt because the corpus we were using was written English, so that they would tend to occur only in the dialogic passages of narrative fiction. So their effect on the total picture was negligible.

(2) We counted all the instances of the operators in their abbreviated form. The main problem here is the *'s,* since this can also be possessive; by our estimate, about two thirds of the occurrences of *'s* (about 70,000, out of just over 100,000) were possessive, leaving about 30,000 as abbreviation of *is* or *has.* The total number of abbreviated operators came to about 80,000. For polarity, these were all to be added to the positive count, For primary tense, about 60,000 were present and rather over 5,000 were past; the remainder were either future or modal, and so outside the population under study.

(3) For clauses with the words *not, no, never* and so on we adopted a strategy that took account of the distinction between *not,* as the usual form of the ordinary finite negative in writing (spoken *they didn't know*, written *they did not know*), and the remaining 18 negative words. Not every occurrence of *not* marks a negative clause, as can be shown by the tag; compare:

Not all the eligible voters turned up to vote, did they? [negative clause]

A not inconsiderable number of voters stayed away, didn't they? [positive clause]

So we devised a strategy for estimating those that were negative (all those where *not* follows a Finite operator, either directly or with one word in between); this yielded about 60,000, or around 70% of all occurrences of *not.* (This was perhaps rather an underestimate.) With the other negative words, we inspected 200 occurrences of each and extrapolated; this had to be done separately for each word, because of the very different proportions involved: for example, whereas 98% of all clauses with *never* turned out to be negative, the proportion was only 70% of those with *no* and only 10% of those with *few*. The total effect of this counting of the negative words was to deduct 70,000 instances from the positive total and add them on to the negative.

(4) We then faced the major problem of counting the instances of simple past and simple present tense forms of lexical verbs. It will be remembered that for the three most common verbs, *be, have* and *do*, these forms had already been counted, because they are the same as forms that occur as operators; and for our purposes we did not need to take account of the distinction between the two. So leaving out these three verbs we picked out from the overall frequency list the first fifty verbs, **as indicated by the ranking of their most frequent form.** (Note that the system counts each form of a word separately, so there is a figure for occurrences of *say*, one for *says*, one for *said* and one for

*saying*, but no overall frequency count, or frequency ranking, for the verb *say* taking all its variants together. English verbs typically have four forms, like *say*; some have five, like *take*; some have three, like *put*; and there is one exceptional verb *be*, which has eight.)

As a rule, the most frequent form of the verb was the base form, e.g. *take*, *put*; in a few cases it was the past form, e.g. *said*. [The figures are given in Table 2.] By inspecting the frequency list we identified the fifty most frequent verbs after *be*, *have* and *do*; this included all verbs of which the most frequently occurring form occurred at least 2,400 times in the corpus of 18 million words, We then split this set of fifty verbs into two sets of 25 each, those above and those below the half-way line; we wanted to examine the two sub-sets separately so that we could see whether the various proportions we were concerned with stayed more or less the same in both halves. The top 25 – the "first swath" – were those whose most frequent form had occurred at least 4,900 times.

Now which forms were we counting? Not v-ng, the *-ing* form, because that never occurs as a finite form. Not v-n, the past participle, for the same reason; but here there is a complication, because for the majority of English verbs the v-n form is the same as the v-d form, that for the past tense, and this we did need to count. Of the other two forms, we wanted all occurrences of v-s, the 3rd person singular present, which is always finite; and of the v-0 or base form, we wanted those instances which are present tense but not those which are infinitive. So the problem of counting them was fairly complicated; and it was complicated still further by the fact that occurrences of some of these forms are often not verbs at all but something else, typically singular or plural forms of nouns. So with words like *look*, *looks*, *turn*, *leaves*, *calls* we have somehow to avoid including the spurious instances into our figures.

For these potentially ambiguous forms we adopted the same strategy as we had done elsewhere: that is, we inspected 200 instances chosen at random, counted all those that were to be discarded, and then extrapolated from this figure on to the total. So for example out of 200 occurrences of *looks*, 40 (i.e. 20%) turned out to be nouns; so from the 1,863 occurrences of the form *looks* we deducted 20% and estimated the total verbal (v-s) instances as 1,490. This of course has the effect of changing the rank ordering of the verbs, and in some cases pushed the verb in question off the bottom of the list altogether (for example, *play* was in the list of first fifty verbs until it turned out that it had got there spuriously, by virtue of its frequency as a noun).

How did we then eliminate the other type of spurious instance, where the word in question was functioning as a verb but was non-finite (v-0 as infinitive, v-d = v-n as past participle)? For each v-0 form (base form), and for each v-d **where this is identical with v-n**, we counted all instances following a Finite operator, either directly or with one word in between (e.g. *might make*, *hasn't made*, *will certainly make*, *is being made*) and discarded them; likewise all **other** instances directly following *to* (*to make*) – that is, other than those that had been discarded already by the first procedure, such as *has to make*.

Finally, and again by a rather complicated procedure, we estimated the number of instances of v-0 that were **imperative**; these we eliminated from the primary tense count. (We left them in the polarity count because they do select independently between positive and negative.) Table 2 gives the figures for the first swath

Table 2. VERBS other than be, have and do: FIRST SWATH, v-0, v-s and v-d forms

| v-0 | NUMBER OF INSTANCES | SPURIOUS INSTANCES OUT OF 200 | CORR-ECTED COUNT | v-s | NUMBER OF INSTANCES | SPURIOUS INSTANCES OUT OF 200 | CORR-ECTED COUNT | v-d | NUMBER OF INSTANCES | SPURIOUS INSTANCES OUT OF 200 | CORR-ECTED COUNT |
|---|---|---|---|---|---|---|---|---|---|---|---|
| say | 12,668 | 1 | 12,605 | says | 6,725 | | 6,725 | said | 47,153 | | 47,153 |
| know | 19,245 | | 19,245 | knows | 2,121 | | 2,121 | knew | 8,016 | | 8,016 |
| get | 18,127 | | 18,127 | gets | 1,554 | | 1,554 | got | 11,939 | | 11,939 |
| see | 17,968 | | 17,968 | sees | 690 | | 690 | saw | 6,705 | 2 | 6,638 |
| make | 14,424 | 0 | 14,424 | makes | 2,986 | 1 | 2,971 | made | 17,093 | | 17,093 |
| go | 16,401 | 3 | 16,155 | goes | 2,377 | 0 | 2,377 | went | 11,677 | | 11,677 |
| come | 14,266 | | 14,266 | comes | 2,909 | | 2,909 | came | 11,928 | | 11,928 |
| think | 13,541 | | 13,541 | thinks | 865 | | 865 | thought | 11,875 | 27 | 10,272 |
| take | 12,517 | | 12,517 | takes | 1,946 | | 1,946 | took | 8,207 | | 8,207 |
| want | 10,287 | 3 | 10,133 | wants | 1,751 | 2 | 1,733 | wanted | 5,734 | | 5,734 |
| look | 9,433 | 30 | 8,018 | looks | 1,863 | 40 | 1,490 | looked | 9,083 | | 9,083 |
| tell | 7,661 | | 7,661 | tells | 680 | | 680 | told | 8,589 | | 8,589 |
| find | 8,181 | 1 | 8,140 | finds | 595 | 9 | 568 | found | 8,212 | | 8,212 |
| ask | 3,745 | | 3,745 | asks | 747 | | 747 | asked | 8,090 | | 8,090 |
| give | 7,876 | | 7,876 | gives | 1,434 | | 1,434 | gave | 4,969 | | 4,969 |
| seem | 3,582 | | 3,582 | seems | 3,791 | | 3,791 | seemed | 7,415 | | 7,415 |
| feel | 6,052 | | 6,052 | feels | 838 | | 838 | felt | 7,111 | 2 | 7,040 |
| become | 6,539 | | 6,539 | becomes | 1,723 | | 1,723 | became | 4,735 | | 4,735 |
| begin | 2,025 | | 2,025 | begins | 860 | | 860 | began | 6,489 | | 6,489 |
| leave | 4,069 | 12 | 3,825 | leaves | 1,554 | 143 | 443 | left | 9,493 | 64 | 6,455 |
| keep | 5,897 | 2 | 5,838 | keeps | 536 | | 536 | kept | 3,660 | | 3,660 |
| turn | 3,952 | 61 | 2,747 | turns | 801 | 36 | 657 | turned | 5,741 | | 5,741 |
| call | 3,817 | 51 | 2,844 | calls | 849 | 79 | 514 | called | 7,012 | 39 | 5,645 |
| put | 11,057 | | *5,432 | puts | 587 | | 587 | put | | | *5,625 |
| use | 7,821 | 100 | 3,911 | uses | 598 | 70 | 389 | used | 8,484 | 84 | 4,921 |
| | | Total: | 227,216 | | | Total: | 39,148 | | | Total | 235,326 |

**Results:** For polarity, all are to be added to "positive" (227,216 + 39,148 + 235,326 = 501,690). For primary tense, v-0 and v-s are to be added to "present" (227,216 + 39,148 = 266,364), and v-d are to be added to "past" (235,326).

"Corrected" scores were derived from "number of instances" scores by the same procedure as in Table 1.

**Note:** *put* is both v-0 and v-d. The totals without *put* were: 221,784 for v-0 and 229,701 for v-d, totalling 451,485. This is a ratio of 0.491 to 0.509. The 11,057 occurrences of *put* were therefore distributed into v-0 and v-d in this same proportion. i.e. 5,432: 5,625.

(from Halliday and James, ibid.)

of 25 verbs, with the totals entered at the bottom. Tables 3 and 4 give the new revised totals for polarity and primary tense. The results now read as follows: for polarity, 1,443,670 instances counted, of which 86.75% were positive and 13.25% were negative; for primary tense, 1,195,710 instances counted, of which 49.55% were present and 50.45% past.

Table 3. Polarity

| POSITIVE | | | | NEGATIVE | | | |
|---|---|---|---|---|---|---|---|
| Finite ops: | modal, and future temporal | 213,573 | | Finite ops: | modal/future | 20,843 | |
| | present temporal | 376,194 | | | present | 26,513 | |
| | past temporal | 371,879 | | | past | 17,035 | |
| | | 961,646 | 961,646 | | | 64,391 | 64,391 |
| Abbr. Finite ops: | modal/future | 15,493 | | Finite op. + *not*: | | 53,995 | |
| | present | 60,893 | | Abbr. Finite op. + *not*: | | 4,779 | |
| | past | 5,443 | | | | 58,774 | 58,774 |
| | | 81,829 | 81,829 | | | | |
| Verbs, Swath 1 (first 25): | v-0 | 227,216 | | Other Negative Clauses: | | 69,507 | 69,507 |
| | v-s | 39,148 | | | | | |
| | v-d | 235,326 | | | | | 192,672 |
| | | 501,690 | 501,690 | | | | |
| | | | 1,545,165 | | | | |
| SUBTRACT: | | | | SUBTRACT: | | | |
| Positive tag: | 1,119 | | | Negative tag: | 1,408 | | |
| Finite op. + *not*: | 53,995 | | | | | | |
| Abbr. Finite op. + *not*: | 4,779 | | | | | | |
| | 59,893 | 59,893 | | | | | |
| Other negative clauses: | 69,507 | 69,507 | | | | | |
| Verbs: | | | | | | | |
| - Finite op. + v-0 | 73,955 | | | | | | |
| - other *to* + v-0 | 58,193 | | | | | | |
| - Finite op. + v-d | 31,211 | | | | | | |
| | 163,359 | 163,359 | | | | | |
| | | 292,759 | | | | | |
| | | | 1,545,165 | | | | 192,672 |
| | | | - 292,759 | | | | - 1,408 |
| | | Total Positive: | 1,252,406 | | | Total Negative: | 191,264 |

**Results:**
Out of a total of 1,443,670 clauses, 1,252,406 were counted as positive and 191,264 were counted as negative. The ratio of positive : negative is therefore 86.75% : 13.25%.
(from Halliday & James, ibid.)

We then used the same strategies over again to deal with the next swath of 25 verbs (see table 5). When we came to inspect these forms for "spurious" instances, an interesting difference appeared: very many more out of this second group had to be disqualified. In other words, within the 25 **most frequent** verbs, most of the forms occur **only** as verbs, whereas in the second swath a large number of the

Table 4. Primary tense

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| PRESENT | | | | PAST | | | |
| Finite ops: | positive | 376,194 | | Finite ops: | pos | 371,879 | |
| | negative | 26,513 | | Finite ops: | neg | 17,035 | |
| | | 402,707 | 402,707 | | | 388,914 | 388,914 |
| Abbr. Finite ops: | | 60,893 | 60,893 | Abbr. Finite ops: | | 5,443 | 5,443 |
| Verbs, Swath 1 | v-0 | 227,216 | | Verbs, Swath 1: | v-d | 235,326 | 235,326 |
| | v-s | 39,148 | | | | | |
| | | 266,364 | 266,364 | | | | |
| | | | 729,964 | | | | 629,683 |
| SUBTRACT: | | | | SUBTRACT: | | | |
| Tag: | 1,476 | 1,476 | | Tag: | 407 | 407 | |
| Finite op. + v-0: | 73,955 | | | Finite op. + v-d | 31,211 | 31,211 | |
| - other *to* + v-0 | 58,193 | | | | | | |
| imperative | 8,817 | | | | | | |
| | 140,965 | 140,965 | | | | | |
| | | 142,441 | | | | 31,618 | |
| | | | 729,964 | | | | 629,683 |
| | | | − 142,441 | | | | − 31,618 |
| | | Total Present: | 587,523 | | | Total Past: | 598,065 |

**Results:**

Out of a total of 1,195,710 clauses, 587,523 were counted as present and 598,065 were counted as past. The ratio of present to past is therefore 49.55% to 50.45%.

(from Halliday & James, ibid.)

forms occurring are words which function **both** as noun **and** as verb. Other than that, the excess of past over present was somewhat less marked in the second swath (56% of past in swath 2, 62% in swath 1, both figures corrected) – though the **totals** are of course very different: about 230,000 for swath 1, under 100,000 (c. 93,000) for swath 2. (Some of the difference is accounted for by the large number of occurrences of the past form *said;* but not all – if all forms of *say* are left out, swath 1 still shows 60% as past.)

We have now taken into account the 53 most frequent verbs in the language: the two swaths of 25 each, plus *be*, *have* and *do*. This has now given us a total of just over 1½ million instances for polarity, just over 1¼ million for primary tense (because of the three categories omitted from the latter count). Note that the second swath of verbs accounted for only about 6% of the total. We would have liked to know how many instances were still unaccounted for – the simple past and simple present instances of verbs having a frequency below about 5,000 (all forms taken together) in the corpus. It turned out that there was no satisfactory way of estimating this figure. I would make a rough guess of somewhere between 100,000 and 150,000; if this is right, then adding them in would not affect the total distribution by more than 1% either way.

The final figures we obtained are given in table 6. They show positive at 87.6% as against negative at 12.4%. They show past and present almost equal, within 1% of 50% each. Let us now look at these in the light of our original hypothesis, which was (polarity) positive 0.9, negative 0.1; (primary tense) past 0.5, present 0.5. There are perhaps three things that could be remarked on.

Table 5. VERBS other than be, have and do: SECOND SWATH, v-0, v-s and v-d forms

| v-0 | COUNT | SPURIOUS INSTANCES OUT OF 200 | CORR-ECTED COUNT | v-s | COUNT | SPURIOUS INSTANCES OUT OF 200 | CORR-ECTED COUNT | v-d | COUNT | SPURIOUS INSTANCES OUT OF 200 | CORR-ECTED COUNT |
|---|---|---|---|---|---|---|---|---|---|---|---|
| need | 7,690 | 74 | 4,845 | needs | 2,726 | 95 | 1,431 | needed | 2,667 | 24 | 2,347 |
| hear | 3,118 | | 3,118 | hears | 164 | | 164 | heard | 4,823 | | 4,823 |
| mean | 4,968 | 7 | 4,794 | means | 4,279 | 97 | 2,204 | meant | 2,448 | | 2,448 |
| help | 6,396 | 60 | 4,477 | helps | 545 | | 545 | helped | 1,349 | | 1,349 |
| believe | 4,444 | | 4,444 | believes | 429 | | 429 | believed | 1,646 | | 1,646 |
| work | 13,630 | 135 | 4,430 | works | 1,498 | 101 | 742 | worked | 2,448 | | 2,448 |
| sit | 1,991 | | 1,991 | sits | 297 | | 297 | sat | 4,094 | | 4,094 |
| remember | 4,088 | | 4,088 | remembers | 188 | | 188 | remembered | 1,417 | | 1,417 |
| bring | 3,049 | | 3,049 | brings | 530 | | 530 | brought | 4,266 | 10 | 4,053 |
| try | 4,222 | 9 | 4,032 | tries | 385 | 17 | 352 | tried | 3,615 | 3 | 3,561 |
| stand | 2,401 | 30 | 2,041 | stands | 700 | 24 | 616 | stood | 3,837 | | 3,837 |
| change | 4,428 | 122 | 1,727 | changes | 1,863 | 163 | 345 | changed | 3,792 | 2 | 3,754 |
| hold | 2,616 | 40 | 2,093 | holds | 508 | | 508 | held | 3,692 | | 3,692 |
| live | 3,892 | 13 | 3,639 | lives | 2,189 | 149 | 558 | lived | 2,396 | | 2,396 |
| talk | 4,168 | 37 | 3,397 | talks | 444 | 120 | 178 | talked | 1,537 | | 1,537 |
| start | 3,334 | 51 | *2,484 | starts | 456 | | 456 | started | 3,396 | 3 | 3,345 |
| let | 6,628 | 15 | *3,257 | lets | 161 | | 161 | let | | | *2,874 |
| understand | 3,193 | | 3,193 | understands | 135 | | 135 | understood | 1,247 | | 1,247 |
| run | 3,795 | 33 | 3,169 | runs | 779 | 46 | 600 | ran | 1,957 | | 1,957 |
| happen | 1,768 | | 1,768 | happens | 1,131 | | 1,131 | happened | 3,153 | | 3,153 |
| lose | 1,361 | | 1,361 | loses | 182 | | 182 | lost | 3,610 | 30 | 3,069 |
| move | 3,011 | 56 | 2,168 | moves | 513 | 82 | 303 | moved | 3,032 | 4 | 2,971 |
| show | 3,847 | 57 | 2,751 | shows | 1,114 | 21 | 997 | showed | 1,784 | | 1,784 |
| set | 6,246 | 40 | *2,655 | sets | 620 | 114 | 267 | set | | | *2,342 |
| read | 4,750 | 5 | *2,460 | reads | 231 | | 231 | read | | | *2,171 |
| | | Total v-0: | 77,431 | | | Total v-s: | 13,550 | | | Total v-d: | 68,315 |

**Results:** For polarity, all are to be added to "positive" (77,431 + 13,550 + 68,315 = 159,296). For primary tense, v-0 and v-s are to be added to "present" (77,431 + 13,550 = 90,981), and v-d are to be added to "past" (68,315).

"Corrected" scores were derived from "number of instances" scores by the same procedure as in Table 1.

**Note:** *let*, *set* and *read* are all both v-0 and v-d. The totals without these three verbs were: v-0, 69,059; v-d, 60,928, which is a ratio of 0.531 : 0.469. The occurrences of *let*, *set* and *read* were therefore distributed into v-0 and v-d in the same proportion, i.e, (let) 6,131→3,257, 2,874; (set) 4,997→2,655, 2,342; (read) 4,631→2,460, 2,171. (from Halliday and James, ibid)

Table 6. Final Totals and Percentages

| Polarity | | | Primary Tense | | |
|---|---|---|---|---|---|
| | Positive | Negative | | Present | Past |
| Table 3 | 1,252,406 | 191,264 | Table 4 | 587,523 | 598,065 |
| Table 5 | 159,296 | | Table 5 | 90,981 | 68,315 |
| | -61,355 | | | -49,458 | -16,516 |
| Totals: | 1,350,347 | 191,264 | | 629,046 | 649,864 |
| Results: | | | | | |
| Positive | 1,350,347 | =87.6% | Present | 629,046 | =49.18% |
| Negative | 191,264 | =12.4% | Past | 649,864 | =50.82% |
| Total: | 1,541,611 | | Total: | 1,278,910 | |

(from Halliday and James, ibid)

The subtraction of 61,355 from column one of Polarity is the result of deducting 23,475 instances of Finite operator + v-0, 21,364 instances of *to* + v-0 (other than the preceding), and 16,516 instances of Finite operator + (v-d = n-n). The subtraction of 49,458 from column one of Primary Tense is the result of deducting 23,475 (Finite op + v-0), 21,364 (other *to* + v-0), and 4,619 instances of v-0 = imperative. The deduction from the final column is of instances of Finite operator + (v-d = v-n). A detailed breakdown of how these figures were arrived at can be found in Halliday & James (ibid) p.62.

As I said in the first part of the paper, the hypothesis was not plucked out of the air; it was based on various bits of counting that I had done, over the years, which although on a very small scale had been carried out carefully and, as far as I was able to do so, accurately and with the same general model of the grammar. So it seems to suggest that counting a sample of 2,000 instances may not be entirely without value for predicting a general pattern. But, of course, it needs to be done again and fully documented.

Secondly, on the other hand, we cannot easily gauge the accuracy of our own figures. We know of numerous places where errors will have been coming in, and there are no doubt other sources of error that we have not yet become aware of. We did our best to reduce the errors; and, when we could think of no way of reducing them further, to make them such as would cancel each other out rather than skew the findings in one direction or the other. Perhaps the most urgent task, at this stage, is to design pattern-matching software that will enable much more accurate estimates to be made.

However, I do not think the figures are so far out as to invalidate the general picture they present; and this leads me to the third point. Of course, it is a coincidence that they are so close to what was predicted; these were idealized modal values, and I am quite sure that such results will not turn up again in any future counting – this was just to encourage us to continue! Do not imagine, by the way, that we could possibly have influenced the results by our choice of strategy; we had no idea how they would develop, and would not have had the slightest notion how to steer them towards any particular outcome. (If one knew enough to do that, one would hardly need to count them). The main significance of the result is that it suggests it would be worth while to count some more.

With the corpus of 200 million words, it would be possible to do this in a way which took account of variation in register, the systematic differences in grammar and lexis that distinguish one text type from another. In the smaller corpus,

information about the source of each text is of course available, but the corpus is not organized on this principle. The larger corpus, the "Bank of English", on the other hand, is stored and indexed for access according to register; so it would be possible to keep an ongoing tally not only of the overall frequencies but of the frequencies displayed by each register that is represented. This would enable us to explore in accurate quantitative terms the notion of register variation defined as the resetting of probabilities in the lexis and in the grammar[7].

As I said at the beginning, corpus linguists often modestly refer to themselves as the data gatherers in the linguistic profession. I do not think data gathering is anything there is reason to be modest about, certainly when it comes to data which are instances of semiotic systems. But in any case, data gathering and theorizing are no longer separate activities (I do not believe they ever were); and whether or not there is any significance in the particular quantitative study reported here, with the potential for quantitative research opened up by corpus linguistics our understanding of language, and hence of semiotic systems in general, seems likely to undergo a qualitative change.

## Notes

1. For the discussion of 'system' and 'structure' as fundamental theoretical concepts, see J R Firth, 'A synopsis of linguistic theory' in Firth et al (1957).
2. See Matthiessen (1983; in press, a).
3. For a general perspective on corpus studies in English, see Svartvik (ed) (1992). For examples of corpus-based lexicogrammatical studies in the 1960s, see Svartvik (1966), Sinclair et al (1970), and Huddleston et al (1970).
4. See Hasan & Cloran (1990) and Hasan (1992).
5. For the "Penman" project in artificial intelligence (text generation), see Mann (1985) and Matthiessen (1985). See also Bateman & Matthiessen (1991).
6. For the COBUILD corpus and the research deriving from it, see Sinclair (ed) (1987).
7. For a detailed discussion of register, including register as a probabilistic concept, see Matthiessen (in press, b).

## References

Bateman, John & Matthiessen, Christian (1991) *Systemic Linguistics and Text Generation: Experiences from Japanese and English*, London & New York: Frances Pinter

Firth, J R (1957) 'A synopsis of linguistic theory' in Firth, J R et al *Studies in Linguistic Analysis*, Oxford: Blackwell (Special Volume of the Philological Society), pp 1-32

Halliday, M A K (1956) 'Grammatical categories in Modern Chinese', *Transactions of the Philological Society*, pp 177-224

Halliday, M A K (1991) 'Towards probabilistic interpretations' in Ventola, Eija (ed) *Functional and Systemic Linguistics: Approaches and Uses*, Berlin & New York: Mouton de Gruyter (Trends in Linguistics Studies and Monographs 55) pp 39-61

Halliday, M A K (1992) 'Language as system and language as instance: the corpus as a theoretical construct' in Svartvik (ed) (1992) pp 61-77

Halliday, M A K & James, Z L (1993) 'A quantitative study of polarity and primary tense in the English finite clause' in Sinclair, John M; Hoey, Michael; & Fox, Gwyneth (eds) *Techniques of Description: Spoken and Written Discourse* London: Routledge, pp 32-66

Hasan, Ruqaiya & Cloran, Carmel (1990) 'A sociolinguistic interpretation of everyday talk between mothers and children' in Halliday, M A K; Gibbons, John; & Nicholas, Howard (eds) *Learning, Keeping and Using Language: Selected Papers from the 8th World Congress of Applied Linguistics, Sydney, 16 - 21 August 1987, Vol. 1*, Amsterdam & New York: Benjamins, pp 67-99

Hasan, Ruqaiya (1992) 'Rationality in everyday talk: from process to system' in Svartvik (ed) (1992), pp 257-307

Huddleston, R D; Hudson, R A; Winter, E O; & Henrici, A (1970) *Sentence and Clause in Scientific English*, London: Communication Research Centre, University College London (for Office of Scientific and Technical Information).

Mann, William C (1985) 'An introduction to the Nigel text generation grammar' in Benson, James D & Greaves, William S (eds) *Systemic Perspectives on Discourse, Vol. 1*, Norwood, N.J.: Ablex, pp 84-95

Matthiessen, Christian (1983) 'Choosing primary tense in English', *Studies in Language* 7.3, pp 369-429

Matthiessen, Christian (1985) 'The systemic framework in text generation' in Benson, James D & Greaves, William S (eds) *Systemic Perspectives on Discourse, Vol. 1*, Norwood, N.J.: Ablex, pp 96-118

Matthiessen, Christian (in press, a) 'Systemic perspectives on tense in English' in Berry, Margaret; Butler, Christopher S; & Fawcett, Robin P (eds) *Grammatical Structure: a Systemic Perspective*, Norwood, N J: Ablex (Meaning and Choice in Language, Vol. 2).

Matthiessen, Christian (in press, b) 'Register in the round: diversity in a unified theory of register analysis' in Ghadessy, Mohsen (ed) *Register Analysis: Theory into Practice*, London & New York: Frances Pinter.

Nesbitt, Christopher & Plum, Guenter (1988) 'Probabilities in a systemic grammar: the clause complex in English' in Fawcett, Robin P & Young, David J (eds) *New Developments in Systemic Linguistics, Vol. 2*, London & New York: Frances Pinter, pp 6-38

Plum, Guenter & Cowling, Ann (1987) 'Social constraints on grammatical variables: tense choice in English' in Steele, Ross & Threadgold, Terry (eds) *Language Topics, Vol. 2*, Amsterdam & Philadelphia: Benjamins, 1987, pp 281-305

Sankoff, David (ed) (1978) *Linguistic Variation: Models and Methods*, N.Y.: Academic Press.

Shannon, Claude E & Weaver, Warren (1949) *The Mathematical Theory of Communication*, Urbana, Ill.: University of Illinois Press.

Sinclair, J M; Daley, R; & Jones, S (1970) *English Lexical Studies*, London: Office of Scientific and Technical Information (Report no. 5060)

Sinclair, John M (ed) (1987) *Looking up: an Account of the COBUILD Project in Lexical Computing*, London: HarperCollins Publishers.

Svartvik, Jan (1966) *On Voice in the English Verb*, The Hague: Mouton (Janua Linguarum Series Practica 63).

Svartvik, Jan (ed) (1992) *Directions in Corpus Linguistics: Proceedings of Nobel Symposium 82, Stockholm, 4 - 8 August 1991*, Berlin: Mouton de Gruyter.

Zipf, G K (1935) *The Psychobiology of Language*, Boston: Houghton Mifflin.

# LOCAL GRAMMARS AND THEIR REPRESENTATION BY FINITE AUTOMATA

*Maurice Gross*
University of Paris 7
Laboratoire d'Automatique Documentaire et Linguistique[1]

In the study of collocations and of frozen sentences (idioms, clichés, collocations, many metaphors and figurative meanings, etc.) one often encounters sets of similar forms that cannot be related by formal rules of either the phrase structure or transformational type. We present examples of such situations and we show how the formalism of finite automata can be used to represent them **in a natural way.**

## 1. Introduction

Transformational grammars are global grammars whose aim is to describe the sentences of a language at a formal level, that is, in strictly combinatorial terms. The descriptions are intended to be complete, namely to attain a coverage as extended as possible of the language. Moreover, the grammar of a language should be such that non-sentences should fall outside of its range in an explicit way. To achieve this goal, Z.S. Harris (e.g. 1964) and N. Chomsky (e.g. 1965) have proposed combinatorial systems specialized in the following way:

-elementary sentences are described by formation rules,

-complex sentences combine sentences (starting from elementary ones) into more complex forms.

Accordingly, transformational rules are of two types:

-unary, transforming an elementary basic sentence into another elementary form,

-binary, transforming a pair of sentences into a more complex one.

Transformations preserve some semantic invariant carried by elementary sentences, implying that related sentences are similar in shape and lexical content. In some cases, a transformational relation is a synonymy relation, but since, for example, it is a transformation that introduces negation it should be clear that a transformational relation preserves the basic meaning but not synonymy. Hence the relation of antonymy preserves the invariant of meaning. In the same way, the oddity of the sentence:

*Your generosity discussed this mandoline*

is preserved in its passive form:

*This mandoline was discussed by your generosity*

In Z.S. Harris' transformational framework, transformations are equivalence relations that operate between two sentences and thus define equivalence classes.

The decision to introduce a given transformation is based on empirical observations subjected to a precisely defined methodology: intuitions about relatedness of sentences, especially intuitions of synonymy of sentences, are numerous, but in order to be formalized into a transformation, such intuitions must be either confirmed by formal arguments that justify a transformational link between them or else left with the intuitive status of synonymous sentences or paraphrases, not representable by grammatical methods. This is the case for distributional relations. Consider a sentence such as:

(1) *Bob worked on this problem*

and other nouns in the complement position:

(2) *Bob worked on this report*

(3) *Bob worked on this question*

Sentences (1) and (3) appear as synonymous, although it is not immediately obvious that the two nouns *problem* and *question* are synonymous, whereas (2) has a different meaning. The relation (1) = (3) is not a transformation. Some synonymy relations will have to be defined independently in order to relate both sentences and exclude the same relation with (2).

In certain situations there are formal reasons to introduce such relations. Consider the two idiomatic sentences:

(4) *This meeting is a pain in the neck*

(5) *This meeting is a pain in the ass*

and the equivalent free sentence:

(6) *This meeting is a pain*

they are clearly synonymous and the two locative noun phrases by which they differ do not contribute to their meaning. Moreover, the parts common to sentences (4) and (5) have identical syntactic and semantic properties, very specific properties since the combination of words is unique. In such circumstances, the argument of idiomatic invariance (J. McCawley 1979, M. Gross 1988) justifies the formal equivalence relation: (4) = (5) = (6), and these three sentences constitute an equivalence class which can be represented by the graph of figure (1).

Such a graph is read from left to right, that is, from the initial state to the final state. Each path represents a sentence. Such finite state graphs or automata are restricted here to finite paths (i.e. no loops are allowed); they are called Directed Acyclic Graphs or DAGs. They represent in a natural way local variants of a given sentence or phrase. Their computational nature makes them efficient in parsing procedure (M. Silberztein 1989).

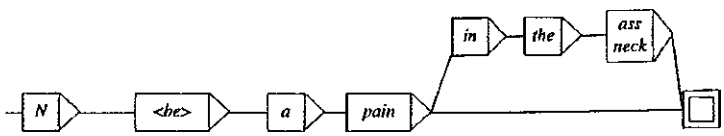



Figure 1

We propose here to describe certain families of utterances by means of finite automata. We will see that such descriptions serve simultaneously several purposes:

- they are special formation rules, they generalize the notion of elementary sentence, and as such they constitute departure points for the application of transformations (unary and binary),
- they constitute a means of representation for equivalence classes built by means of transformations (E. Roche 1992).

At this point, we must make more explicit our representation of sentences in order to define clearly the new relation we have just introduced in the formation component of elementary sentences.

Our topic is that of elementary sentences, that is, sentences of the shape subject-verb- essential complements (if any). We write $N_0$ *V W* for such general shapes. Since each verb governs a specific string of complements, we will specify structures in the following way:

$N_0$ *V W* =: $N_0$ *give* $N_1$ *to* $N_2$

The symbols $N_i$ (i = 0, 1,...) stand for noun phrases.

The question of the attachment of a preposition (here *to*) to its noun phrase is open. Indeed, some device will have to be added to such representations, in order for example to distinguish English from French where preposition stranding occurs much less often.

The complement noun phrases $N_1$ and $N_2$ are numerically indexed in order to precisely define the combinatorial effect of transformations. For example, we write:

$N_0$ *give* $N_1$ *to* $N_2$ = $N_0$ *give to* $N_2$ $N_1$

for the stylistic transformation that tends to order both complements according to their length:

*Bob gave the book to one of the students*
*Bob gave to the student the book he went through during the semester*

The transformation:

$N_0$ *give* $N_1$ *to* $N_2$ = $N_0$ *give* $N_2$ $N_1$

is different, since it changes the status of the second complement, which may undergo Passive:

*Bob gave the student the book he went through during the semester*
= *The student was given the book Bob went through during the semester*

In order to indicate this change of function, we can write:

$N_0$ *give* $(N_2)_1$ $(N_1)_2$

meaning that the second complement has become first (authorizing passivization) whereas the direct object has become second complement. The stylistic length permutation cannot apply to this form, as in:

*?*Bob gave the book the student who has asked so many questions about its author*

By specifying information in this way, we define complete classes of equivalence on a formal basis, such as:

*Bob described the volcano*
*Bob did not describe the volcano*

*Bob is a describer of the volcano*
*Bob made a description of the volcano*
*The volcano has been described*
*The volcano is not describable*
*The volcano is undescribable*
*The volcano has a certain description by Bob*
*The volcano has no description*
*The volcano is without description*, etc.

It is important to stress the use of Nominalization and Adjectivization relations. By introducing the notion of support verb (*to be*, *to have*, etc., Z.S Harris 1964, M. Gross 1981), we account for the full derivational morphology of the verb by syntactic methods. In this way, we eliminate the so-called morphological level from the synchronic description.

## 2. Examples of local grammars

### *2.1 Example 1*

Let us now consider another example of the equivalence relation between sentences at the formation level. The following idiomatic expressions are synonymous:

*Bob lost his cool*
*Bob lost his temper*
*Bob lost his cork*
*Bob lost his self-control*
*Bob blew a fuse*
*Bob blew a gasket*

The direct complement is frozen; namely the determiners are frozen (e.g. the possessive adjectives are obligatorily coreferent to the subject, and no modifier is allowed for the complement nouns). We observe variants for these forms, but they have the same meaning:

*Bob blew his cool*
*Bob blew his temper (up)*
*Bob blew his top*
*Bob blew his cork*
*Bob blew his stack*

Since these sentences share many features, we will represent their similarities, which leads us to structure this list of sentences into a **local grammar** for these utterances.

First we can factor out the sequence[2]:

$N_0$ (=:*Bob*) *lose* $Poss^0$

which is shared by sentences differing only in respect of the frozen noun complement.

We can proceed in the same way with the verb *to blow*. However, the determiners are more varied; we have two new cases: *a* (*fuse* + *gasket*) and *the* (*lid*).

If we compare the complements of both verbs *lose* and *blow*, we find a common set of three nouns with $Poss^0$. The noun *stack* must however be left out of this group:

**Bob lost his stack*

This set of similarities and differences is represented in the graph of figure 2.

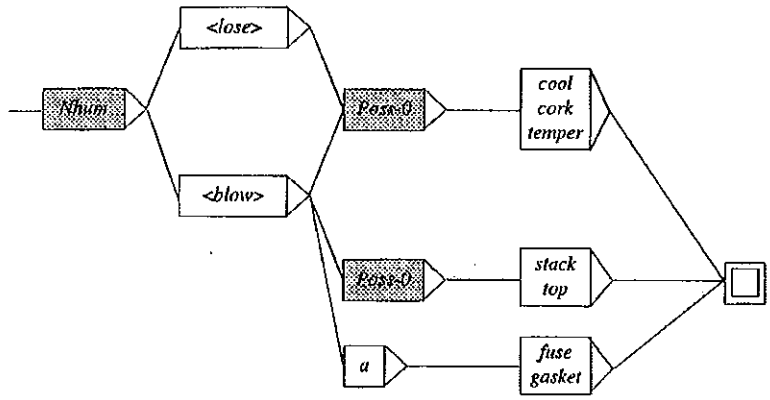



Figure 2

But we can refine the description and extend it. First, the nouns are not all frozen to the same extent; in other words the combinations *Verb-Noun* are not all idiomatic in the same way. Thus the noun *temper* has a range that goes beyond the two verbs discussed, but it does not have the full autonomy of the synonymous noun *self-control*:

*Bob lost his self-control*
*Bob lost the remarkable self-control he has always displayed*
**Bob lost the remarkable temper he has always displayed*

Other modifiers are common to these two nouns, but excluded for others:

*Bob lost his proverbial self-control*
*Bob lost his proverbial temper*
**Bob lost his proverbial cork*

The nouns *cork*, *fuse*, *gasket*, *stack*, *top* do not appear to be used in the same idiomatic way outside of the sentences of figure 2, but this is not the case for *temper* and *cool*. We observe:

*Bob (kept + controlled + held) his temper*
*Bob (is in + is out of) temper*
*Bob (got + flew) into a temper*
*Bob kept (E + his) cool*

All of these sentences contain a support verb, and we observe here common restrictions on the combinations between support verbs and their supported nouns. Moreover, we observe new frozen forms:

*Bob flipped his lid*

and synonymous forms with similar structures:

*Bob (was in + flew into) a rage.*

In order to include these sentences in the grammar of figure 2, one has to defactorize some of the paths:

- we can isolate *temper*, in order to add the other support verbs, and authorize *Modifiers*,

- we have to isolate *cool* in order to introduce the support verb *to keep*.

The resulting grammar is shown in figure 3.

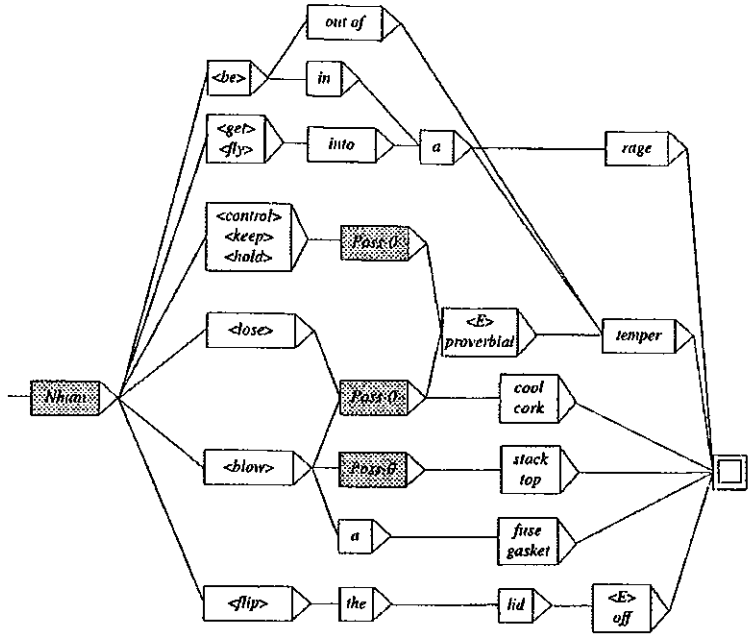



Figure 3

*2.2 Example 2*

Let us now discuss a different example: adverbial phrases that correspond to dates.

The graph **DateRounded** of figure 4 represents a family of date forms that are semantically similar, in the sense that the years are rounded to the nearest tens. The tens are excluded. These figures are given in the two forms: alphabetical and numerical.

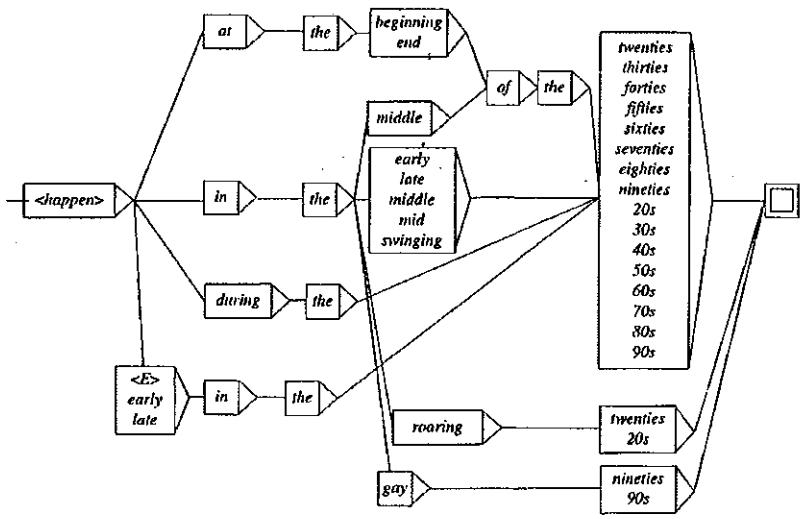



Figure 4

The range of ten is loosely divided into three periods by the terms *beginning*, *mid*, *middle* and *late*; the adjective *swinging* belongs to another semantic register. We have added two idiomatic phrases.

Prepositional variants are represented.

The positions of the adjectives *early* and *late* are represented by two different, hence independent, paths, hiding a possible syntactic relation representable by the permutation rule:

*in the (early + late) sixties*

= *(early + late) in the sixties*

*2.3 Example 3*

Let us undertake a description of more precise dates[3], namely forms that occur in sentences such as:

(1) *The incident took place on Tuesday May 2nd, 1969*

There are variants for this date form.

The name of the day is redundant; it can be derived from the numerical date by using a calendar. Hence the date in (1) is equivalent to the abbreviated form:

(2) *on May 2nd, 1969.*

In a similar way we can argue that the numerical name of the day can be reconstructed from the expression:

(3) *on the first Tuesday of May 1969*

by using the arithmetical condition: *2 is smaller than 7*, but in a first stage of description we will not attempt to describe (3) as a formal variant of (2).

It is interesting to observe that the phrase of date is naturally described as a prepositional noun phrase, since it has the properties of this grammatical notion. But the internal structure of this phrase cannot be analyzed in traditional terms:

- *Tuesday* and *May* seem to be nouns, perhaps proper names, but of such a specific nature that no terminology has been made up for them;
- *2nd* and *1969* are numerals; *2nd* is an *ordinal number* but is *1969* a *cardinal number*?

Moreover these numerals qualify 'nouns', but the nature of the relation between *2nd* or *1969* and the names of months is extremely specific and terms like 'adjective' or 'determiner' are not relevant to their description. Meanwhile, the pattern of dependencies between all these lexical elements is strictly defined and obeys regular rules that we are going to state:

- in the position of *Tuesday* any of the seven names of day may occur,
- in the position of *May*, any of the twelve names of month can occur,
- in the position of *2nd*, we find different forms:

*1st, 2nd, 3rd, 4th, 5th, ..., 21st, 22nd, ..., 30th, 31st*

and we observe numerals from *1* to *31* without the mark of order. In the position of *1969*, a whole range of numerals is allowed. Numerals of years are specific but depending on the calendar or on the range of time (historical times, geological

times) specific modifiers may have to be appended to them (e.g. 800 AD, 400 B.C.). We will use the symbol *NumYear* to represent this set of numerals.

The name of the day is entirely optional and, as mentioned above, its omission does not change the meaning of the phrase. Depending on the context of the sentences that include dates, other parts of (1)-(2)-(3) can be omitted. Hence, the following form is accepted:

*The incident took place on May 2nd*

implying that the year is either the year when the sentence was uttered or a year already mentioned in the context. Hence omitting *NumYear* is exactly like omitting a pronoun; a coreference effect is created and all the problems of attaching the truncated date to a full date are the general problems of the location of an antecedent for a pronoun.

The following forms of date are also possible:

(4) *on Tuesday May the 2nd of 1969*

(5) *on Tuesday the 2nd of May of 1969*

The distributions of names of day and month and *NumYear* are identical; the mark of order must be attached to the numeral of day. The name of the day and *NumYear* can be omitted exactly as in (3). The shape (4) can be considered as related to (3) by the omission of the determiner *the*. The shape (5) presents a different word order that could perhaps be related to that of (4) by a transformation. Nonetheless we will describe the sequences of type (5) independently from those of type (4).

In (4) and (5), the name of the month can be omitted, and the context should allow for the interpretation of the truncated form:

(6) *The incident took place on the 2nd*

When the month is omitted the day is allowed but *NumYear* is forbidden:

*The incident took place on Tuesday the 2nd*
**The incident took place on the 2nd 1969*

Notice that (6) can be seen as the result of the omission of *May* in (4) or of *May* in (5). We will choose the solution: omission in (5), for the purely formal reason that *May* is contiguous to *1969*, making the dependency between both abbreviations easier to state, as will be seen below.

Omitting the numeral of day is not possible:

**on May 1969*

but we do have the form:

(7) *in May 1969*

where *NumYear* can be omitted: *in May*.

Attempting to include (7) in our paradigm of dates involves substituting *in* for *on*. However, unlike the other substitutions we considered, this substitution is restricted to a particular substring of the departure string; hence, on this basis only, we see no benefit in including the forms (7) in the grammar of dates we can now write. On the other hand, the situation is similar with the form:

(8) *on Monday*

Since in (8) the name of the day is obligatory, whereas it was optional in all other forms, and since *NumYear* is forbidden whereas optional elsewhere, the question arises whether to describe it as a separate form of date.

We mentioned that *NumYear* needed further description; this description could take this same form, and the corresponding automaton could be appended to the automaton of dates.

We could introduce further details into the date by adding the time of the day:

*on Monday, May 2nd, 1969* *at noon*
*at four o'clock*
*at 4 p.m*
*at 16 hours and 32 minutes*

The utterances found to the right of *at* can be described exactly by the same method; the construction of the automaton will raise exactly the same problems of substitution, abbreviations and compatibilities between strings and substrings.

Again the resulting automaton could be optionally appended to the right of the final state of the automaton of dates given in figure 5.

## 3. Transformations of finite state grammars

So far we have dealt with a family of strings formally defined by two operations on a given sequence of words:

- allowing **substitutions** of words,
- allowing **omission** of words[4].

The various examples of descriptions we have just given are satisfactory only to a certain point. We have been mainly using a formal principle of factorization of the word sequences that are common to several utterances. We have been forced to repeat some sequences within the same graph (i.e. within the given local grammar); in other terms, the principle fails in such cases.

In one case at least, the source of the failure is clear, and we already mentioned that the permutation rule:

*(early + late) in the sixties = in the (early + late) sixties*

could save a subgraph in figure 4. In a more general way, we face a broad limitation: permutation rules cannot be handled in a natural way by finite-state grammars.

This observation holds for the grammar of figure 5: various boxes (subgraphs) have been duplicated:

- two for the names of day and of month,
- two boxes for *NumYear*,
- three boxes for the numerals of day.

Can we save such duplications by introducing permutation rules, that is transformational rules? The answer is clearly no in the case of the date adverbials starting with *in*, since no numeral of day is allowed.

If we examine the two subgraphs that include the names of day, we immediately see that the determiner *a* is in complementary distribution with the *Modifiers* of the

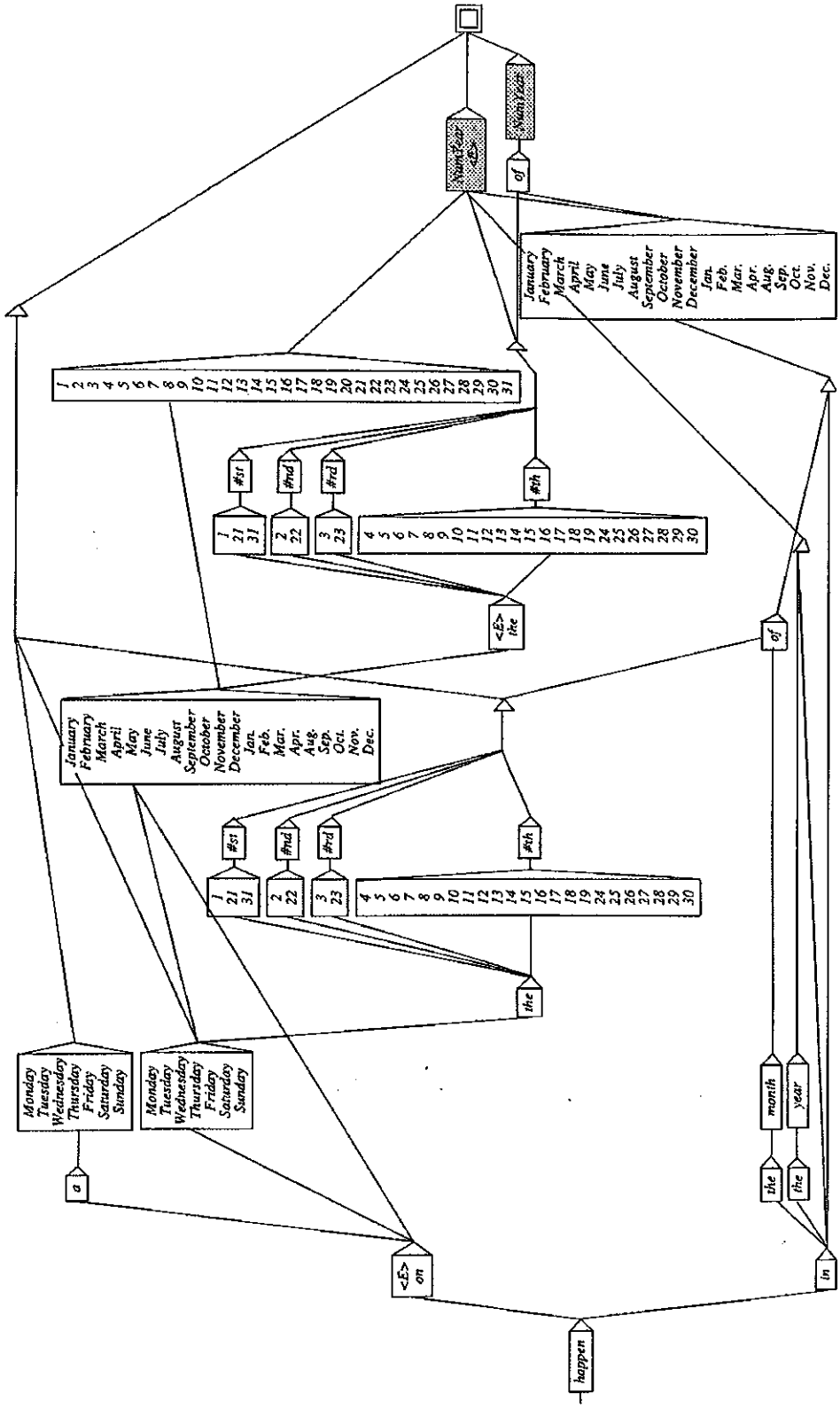

Number
Month
of
January February March April May June July August September October November December Jan. Feb. Mar. Apr. Aug. Sep. Oct. Nov. Dec.
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30 31
#st
#nd
#rd
#th
1 21 31
2 22
3 23
4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 24 25 26 27 28 29 30
<E> the
January February March April May June July August September October November December Jan. Feb. Mar. Apr. Aug. Sep. Oct. Nov. Dec.
#st
#nd
#rd
#th
1 21 31
2 22
3 23
4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 24 25 26 27 28 29 30
the
Monday Tuesday Wednesday Thursday Friday Saturday Sunday
Monday Tuesday Wednesday Thursday Friday Saturday Sunday
a
<E> on
happen
of
month
year
the
the
in

Figure 5

names of day placed to their right. This type of complementation is linguistically significant and found in other contexts. Hence, some equivalence rule such as:

*a = Modifier (=:June 6,1969)*

holds, up to a permutation rule, since we have:

*a Monday* and *Monday June 6,1969*

Thus, introducing such transformations would save the duplication of the names of day. In the same way, a transformation that has the following effect:

*on the 6th of May = on May 6*

could save duplications of subgraphs.

On the whole, from the point of view of savings that we have developed, a notion of non-redundant grammar can be outlined as a system of two components:

- formation rules constituted by finite-state graphs generating elementary sentences and phrases,
- transformation rules that modify the initial graphs, introducing variants, mainly variants of word order.

A typical and general example is that of adverbial permutations. Consider the sentence:

*People lost their cool*

and the adverbial phrase *in the sixties*. The following combinations of these two utterances are accepted:

*In the sixties, people lost their cool*
*People, in the sixties, lost their cool*
*?People lost, in the sixties, their cool*
*People lost their cool in the sixties*

Such sentences are easily described in a general way by specifying their constituent structure:

$N_0VN_1$

and by stating that the adverbial complement can occur at any constituent boundary of this structure.

Such a formulation of the rule cannot be made directly on the strings defined by our finite automata because there is no indication of constituent boundaries: transitions between states, that is word boundaries, are all of the same nature.

Various formal solutions for handling this situation are possible. For example, we can modify figure 2 by indicating the places where the adverbial strings may occur. We present such a modified grammar[5] in figure 6.

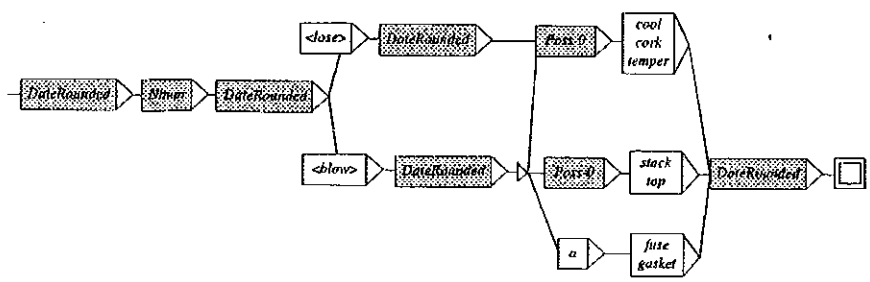



Figure 6

However, the grammar of figure 6 has a standard interpretation: in a shaded box the name of an automaton is indicated which is to be inserted in the automaton of the sentences and which is of the same nature as this main automaton[6]. For example, an adverbial such as *in the sixties* appears four times in the sentences of the corresponding grammar, thus generating unacceptable sentences. Our solution does not allow for restrictions of the form:

(R) *Date Rounded* can occur only once in the grammar.

Hence, a special device must be added to the finite-state formalism. The device (R) is equivalent to a transformation that would move the adverbial to any of its indicated positions.

To sum up these observations and proposals, we have to modify the classical component of formation rules in the following way:

- finite state grammars are used to generate sentences as strings of words,
- a constituent analysis must be superimposed on these strings.
- then the transformational component must be generalized so as to operate on finite-state graphs with marked constituent structure.

Taking into account the fact that lexically frozen structures are more numerous than free ones, the generalization we propose should substantially modify the shape and scope of current parsers.

## Notes

1. Institut Blaise Pascal, CNRS. I am indebted to M.Salkoff for substantial improvements.
2. *Poss*$^0$ represents possessive adjectives coreferent to the subject $N_0$.
3. For detailed descriptions of dates in French, see M. Gross (1990), D.Maurel (1990).
4. We could consider omission as substitution of the null word for a given word.
5. For the sake of simplicity we did not do it for the more complete grammar of figure 3.
6. In fact, the system of programs *InTex* (M.Silberztein 1993) compiles automatically the complete grammar (made of the main sentences and the adverbials) into a recognition procedure that locates in texts each sentence generated by the grammar.

## References

Gross, Maurice (1981). Les bases empiriques de la notion de prédicat sémantique, *Formes syntaxiques et prédicats sémantiques*, A. Guillet et C. Leclère eds., *Langages* N° 63, Paris: Larousse, 7-52.

Gross, Maurice (1988). Methods and Tactics in the Construction of a Lexicon-Grammar. In *Linguistics in the Morning Calm*, Selected Papers from SICOL 1986, Seoul: Hanshin Pub. Co., 177-197.

Gross, Maurice (1990). *Grammaire transformationnelle du français. 3-Syntaxe de l'adverbe*, Paris: ASSTRIL, 670 p.

Harris, Zellig (1964). Elementary Transformations, Philadelphie: University of Pennsylvania, *TDAP* N° 54. Reprinted in *Papers in Structural and Transformational Linguistics*, 1970, Dordrecht: Reidel.

Maurel, Denis (1990). Adverbes de date: étude préliminaire à leur traitement automatique, *Lingvisticae Investigationes*, XIV:1, Amsterdam-Philadelphia: John Benjamins, 31-63.

McCawley, James (1979). *Adverbs, Vowels and Other Wonders*, Chicago: University of Chicago Press, 84-95.

Roche, Emmanuel (1992). Une répresentation par automate fini des textes et des propriétés transformationnelles des verbes, *Lingvisticae Investigationes*, XVI:2, Amsterdam-Philadelphia: John Benjamins.

Silberztein, Max (1989). *Dictionnaires électroniques et reconnaissance lexicale automatique*, Thèse de doctorat, Université Paris 7: LADL.

Silberztein, Max (1993). *Dictionnaires électroniques et analyse automatique de textes: le système INTEX*, Paris: Masson.

# "SWEETLY OBLIVIOUS": SOME ASPECTS OF ADVERB-ADJECTIVE COMBINATIONS IN PRESENT-DAY ENGLISH[1]

*Stig Johansson*
University of Oslo

## 1. Introduction

One of the most remarkable characteristics of language is the ease with which the same word or pattern can be put to different uses, without seeming to cause problems of communication. A *sad person* is sad, a *sad story* makes you sad. A *candid confession* is candid, a *candid recording* provides candid pictures but the recording event is rather the reverse of candid. *Fire protection* means protection against fires, *forest protection* is naturally interpreted as protection of forests. Examples can be multiplied of polysemy in lexis, syntactic ambiguity, and diversity of communicative functions of sentences in context.

Combinations of adverbs and adjectives seem straightforward. Surely this is just a case of modification by degree, as shown by examples like: *absolutely certain, terribly bad, totally wrong*? But it is easy to find examples which cannot be fitted into this simple pattern. Here belongs the example in the title of this paper, *sweetly oblivious*, as well as a host of other expressions: *blissfully ignorant, boyishly handsome, sadly inefficient*, etc. So what are the characteristics of adverb plus adjective combinations?

## 2. Material

The study reported in this paper is based on a collection of examples from the tagged LOB Corpus (cf. Johansson *et al.* 1986). All sequences of a word tagged RB (adverb) followed by a word tagged JJ (adjective) were retrieved from the corpus[2]. This yielded a collection of some 4,000 examples. Not all the examples retrieved in this manner are relevant for our study of adverb-adjective combinations. As the corpus is not syntactically parsed, there is no way of distinguishing sequences where the adverb modifies the adjective (as in *absolutely true*) from those where the adverb functions as an adverbial on the clause level (as in *it is probably true*). The former type is the more frequent one and provides the basis for the discussion below.

The combinations were sorted both by the adverb and by the adjective. This made it easy to find all the adjectives preceded by a particular adverb and, conversely, all

the adverbs preceding a particular adjective. Some examples follow where a combination is found more than once; the frequency is given within parentheses:

| | |
|---|---|
| *absolutely* – | beastly, certain (2), clear (2), complete, delighted, fine, first-class, full-fruit, futile, good, honest, impossible, independent, indispensable, motionless, necessary, right, rotten, safe, simultaneous, standard, still, unfettered, wonderful, wretched, wrong |
| *totally* – | different (3), false, inadequate, real, unfit |
| *utterly* – | alien (3), confounded, degenerate, different (2), extraordinary, fearless, reliable, undistinguished, unforgivable, unreadable, unsuitable, weary |
| – *different* | clearly, completely (4), entirely (3), essentially, fundamentally (2), greatly, little (2), morphologically, naturally, no (3), numerically, quite (8), rather, refreshingly, significantly (2), slightly (5), somewhat (2), strangely, subtly, totally (3), utterly (2), vastly (3), wholly, widely |
| – *right* | absolutely, dam, entirely, exactly, just (5), quite (8) |
| – *wrong* | absolutely, all, completely, demonstrably, hopelessly, patently, really, spectacularly |

Just a glance at the combinations with *different* reveals that there is no single semantic pattern, and it is this variety of semantic patterning which we will now focus on.

## 3. Semantic patterns

Previous discussions of the semantics of adverb-adjective combinations have focused on the expression of degree, with only occasional references to other semantic patterns. Here belongs the work of Altenberg (1991), Bäcklund (1973), Borst (1902), Jonson (1967), Kirchner (1955), Spitzbardt (1954, 1965), etc. Quirk *et al.* (1985: 445-448) draw attention to other patterns, as does the *Collins COBUILD English Grammar* (Sinclair *et al.* 1990:93-97), and some interesting observations are found in Allerton (1987:16-18), Poldauf (1959) and Swan (1988:21ff.). As far as I am aware, there is no description which is sufficiently delicate to account for the variety found in the LOB Corpus material.

It will quickly become apparent that the task is far from straightforward. The classification certainly cannot be exhaustive. There is a great deal of overlap, and there are examples which do not fit neatly into any one pattern. For this reason, there will be a great deal of exemplification, except for the well-known category of degree expressions[3].

### *3.1. Degree and extent*

The number of degree adverbs is very large. Spitzbardt (1965:353) reports that he registered about 700 adverbs of degree in his dissertation (Spitzbardt 1954), based on a study of more than 5,000 pages of Modern English text, and he adds that, together with material recorded in other studies, 'we arrive at a sum total of about 1,000 intensive and restrictive adverbs of degree in English'.

What is more interesting than the number, which is certainly not fixed[4], is the range and variety of degree expressions. Spitzbardt (1965:355) divides adverbs

of degree into nineteen semantic fields representing two main spheres, a 'predominantly objective-gradational sphere' (*largely, lightly, completely*, etc.) and a 'predominantly subjective-emotional sphere' (*attractively, exquisitely*, etc. expressing a positive evaluation; *badly, furiously*, etc. expressing a negative evaluation). More recently, Quirk *et al.* (1985:445) have drawn a distinction between amplifiers, which 'scale upwards from an assumed norm' (*awfully, downright*, etc.), and downtoners, which 'have a generally lowering effect' (*almost, fairly*, etc.).

A type of degree expression which is not focused on elsewhere is found in examples like:

brilliantly clever G36:47
broadly general G58:6
clearly evident J46:147
flatly dull G43:106
furiously indignant K13:141
identically similar N04:5
monotonously uneventful G44:25
vitally important E37:24, K05:43
vitally necessary A27:21, J43:65

These expressions are tautological, but have an intensifying effect. We also find the opposite type of example, where contradictory terms are combined (oxymoron):

discreetly indiscreet G13:146
falsely true (not part of the LOB Corpus material)

The adverb describes the manner in which the quality is manifested (cf.3.3). There is of course also a lowering effect.

Adverb-adjective combinations which express more than just degree or extent are abundant in the LOB Corpus material. Very often there is an element of value judgement (3.9), corresponding to Spitzbardt's 'subjective-emotional sphere'.

### *3.2. Emphasis*

Quirk *et al.* (1985:447) take up a class of adjective modifiers called 'emphasizers', which 'add to the force (as distinct from the degree) of the adjective': *really beautiful, just impossible*, etc. Such examples are not uncommon in the LOB Corpus material:

clearly different G63:108
definitely helpful G47:157
exactly alike E23:24
just right F03:122
positively embarrassing G39:94
really important A40:138
strictly logical J53:80
truly disinterested C12:162

As Quirk *et al.* (*loc.cit.*) point out, the effect is often similar to that of degree expressions.

*3.3. Manner*

Manner is a notion frequently expressed by adverbials on the clause level. It is also commonly found in adverb-adjective combinations:

a few *accidentally conspicuous* individuals G13:75
fastidiously calculated and yet *agreeably spontaneous* C01:80
afraid of her husband and yet *arrogantly proud* N25:176
the river runs *artificially straight* G17:125
the model, although *attractively simple* J46:9
so daring and yet so *audaciously tempting* C10:85
the Tyn Church, *austerely Gothic* G66:15
*boyishly handsome* at other times P03:107
... said Lord Undertone, *carefully casual* M03:15
the affluent middle class who now had plenty of lovely-ugly to be *coldly elegant* in R03:57
a *comfortingly familiar* form G51:65
the father-in-law, always *comically grotesque* C03:189
"concerned" in the sense of deeply and *compassionately aware* G60:134
the *consistently favourable* attitude F27:146
clarifying and *enjoyably articulate* J60:57
a *falsely high* value F34:174
*innocently negligent* ... shareholders A25:84
a *nervously conscientious* child G22:188
in this *oddly unbiblical* way D10:2
they are *openly critical* about the higher landing fees A15:73
all situations except the *overtly sexual* G77:120
a *quietly effective* production C03:130
had every intention of staying *solitarily true* to Peggy P20:152
a *splendidly efficient* play C03:136
the end-product does not become *tiresomely assertive* F21:130
being *uniformly brown* in colour J03:127

Examples can easily be multiplied. Allerton (1987:17) points out that manner expressions may border on degree: *cautiously optimistic*, *easily accessible* (cf. *slightly optimistic*, *very accessible*). More important, they frequently also express a value judgement (cf.3.9).

*3.4. Time*

Adverbs modifying adjectives may express time, as in the following examples:

his *always bright* turn of speed A07:50
the *ever delightful* blacksmith G22:127
their *formerly essential* skill F36:120
cursory and *frequently inaccurate* oral reading H03:178
a *long overdue* reform measure J43:65
*newly independent* churches B03:13
his *often dangerous and adventurous* life B08:150
its *once formal* gardens L09:53
*rapidly fatal* heart failure J16:9
a long and *sometimes rambling* story C01:17

the soft, *still dew-moist* grass l03:156
the man's *suddenly limp* hand L03:9
the *usually calm, reserved* Kay P23:153

Note that *sometimes* in the following example (in contrast to the one above) does not really express time:

this gently undulating, *sometimes flat* surface J03:58

What is expressed here is space rather than time, which brings us to our next category.

*3.5. Space*

Only occasionally do we find adverb-adjective combinations where the adverb expresses a spatial notion:

an *internationally famous* restaurateur E19:146
*locally resident* civilians J30:35
Mr Powell, white-faced and *outwardly respectable* G27:142
the *universally unpopular* Scots C12:141
such tests became *widely acceptable* J36:160

Admittedly, *outwardly* could also be placed in the category of 'respect', to be dealt with below (3.6). *Internationally* and *widely* border on expressions of degree/ extent (3.1); note that *widely* is clearly a degree expression in examples like these from the LOB Corpus: *widely different, widely variant.*

*3.6. Viewpoint and respect*

Many adverbs express the respect in which a quality is applicable or the point of view from which the quality is viewed. Examples are:

academically barren, administratively possible, aesthetically valid, analytically prior, commercially unrealistic, economically disastrous, emotionally offensive, financially secure, functionally independent, grammatically equivalent, intellectually alert, logically predictable, mentally defective, numerically different, physically strenuous, politically relevant, psychologically asleep, socially superior, statistically significant, structurally separate, tactically useful, technically astute, theoretically possible

Quirk *et al.* (1985:448) place such examples in their 'viewpoint' category. Allerton (1987:17) uses the term 'aspect' in much the same way, which is less apt as aspect in linguistic discussion normally refers to something quite different. The two terms in the heading above both seem applicable, and no attempt will be made to distinguish between them.

Whatever terms are used, we should note the relationship to other semantic patterns. Allerton (1987:17) gives *sexually attractive* as an example of his 'aspect' category but points out that it is not far from the manner type. The relationship is especially close to the categories taken up next, where the role of the observer is even more prominent.

*3.7. Evaluation of truth*

This category corresponds to truth-evaluating disjuncts (cf. Quirk *et al.* 1985:62f.) on the clause level[5]. Expressions of this type include:

*admittedly extensive* rent control E28:34
her *apparently blameless* past C09:194
war is the *most obviously insane* G13:70
*ostensibly minor* vexations F21:155
a different and *perhaps better* use H05:146
this other man creature is *plainly useless* M06:70
playful, and even *possibly lascivious* hours lay before me K26:11
the small *presumably private* and struggling shops E22:138
dust from *probably Dickensian* times R03:89
*seemingly authoritative* reports A21:121
three *supposedly equal* areas J28:94

The adverbs express notions like conviction (e.g. *admittedly*), doubt (e.g. *perhaps*), and distance from reality (e.g. *apparently*).

*3.8. Basic and typical qualities*

Another sort of evaluation is expressed in examples like:

a *basically evil* policy B01:174
an *essentially creative* process C13:21
*fundamentally Italian* operas G44:55
*inherently progressive* subjects G64:92
the good influence of *naturally anxious* wives B18:173
a *typically British* feature B16:126

The adverbs are used here to underline that the qualities apply in principle or are seen as basic or typical characteristics. The category is related both to the preceding one (Quirk *et al.* 1985:621 do indeed include expressions with *basically, essentially* and *fundamentally* in their category of disjuncts expressing degree of truth) and to the one taken up next. In all these cases we are concerned with some kind of evaluation.

*3.9. Value judgement*

Expressions of degree (3.1) and manner (3.3) frequently convey a value judgement. Some more examples of combinations expressing a value judgement are:

its *absurdly long* wing-filaments C11:221
this loss is *acceptably small* J04:165
he has been *admirably thorough* C08:33
it must have seemed *excitingly new*, even revolutionary C02:140
it too often rings *frighteningly true* C11:184
her new and refurbished hotels and restaurants are *refreshingly different* E22:133
the *sadly inexperienced* Miss Pendleton K19:152
it was *sickeningly clear* to Rachel that ... P27:141
she would show Adrian that he was wrong, and *stupidly old-fashioned* at that P08:101
wages paid in the manufacture of motor vehicles are *unfairly high* J45:52

These combinations admit a paraphrase with 'so': 'so long that it is absurd', 'so small that it is acceptable', etc. Such examples are extremely frequent.

The judgement may have nothing at all to do with degree or manner, as in:

contest the *naturally considerable* bills G54:202

The judgement here could be paraphrased as 'in accordance with expectations'. The following example is more unusual:

Colmore thought of his own parents, now *safely dead*: his mother's wen, his father's lack of aspirates. With such a background one could never be really safe however brilliant one was. K01:29

The adverb-adjective combination compresses the meaning 'I am safe now that my parents are dead'. Such compression of meaning is found both in adverb-adjective combinations and with adverbials on the clause level (see Johansson and Lysvåg 1987:258, 260). Additional examples are given in the next section.

*3.10. Quality and state*

At the outset, it seems unlikely that adverbs should be used to express qualities and states. This is surely the province of adjectives. But we find clear examples of adverbs used in this manner outside adverb-adjective combinations (the second example is quoted from Johansson and Lysvåg 1987:260):

Henriette saw the weaving figure of an Apache warrior reel *nakedly* on a pony N20:192

She wanted to be *warmly* at home. ('be warm and be at home')

Now consider these examples of adverb-adjective combinations:

Mary kept her voice *calmly reasonable* L05:205
one *cheerfully scandalous* anecdote L06:156
his *genially informal* manner F23:64
his look, always *gravely compassionate* L10:193
my notices were *generously kind* about him G28:139
*harmlessly facetious* remarks C09:162
his wife, always *palely appealing* C03:189
the streets are *tranquilly sunny* and still C06:7

The adverbs in most of these examples could be regarded as expressing manner (3.3), but they are not far in effect from coordination: *calm and reasonable, cheerful and scandalous, genial and informal, grave and compassionate, generous and kind, harmless and facetious, pale and appealing*. The last example, which is admittedly literary in flavour and a bit unusual, is best interpreted in this way, as degree, manner, and the other major patterns are excluded. Note that the quality of tranquillity is also expressed by explicit coordination with *still*. The effect is a foregrounding of the quality of tranquillity.

Another example which does not fit into the major categories dealt with before is the expression in the title of this paper:

Only the Labour Party remains *sweetly oblivious*. G67:193

To this could be added similar examples with *blissfully*: *blissfully ignorant, blissfully unaware* (only the latter evidenced in the corpus). Here the adverbs seem to express a state, reflecting corresponding expressions with adjectives and nouns: *oblivion is sweet / sweet oblivion, ignorance is blissful / blissful ignorance*. In effect, the writer conveys an ironic comment. In other words, the apparent simplicity of expression compresses a great deal of meaning.

To further illustrate the compression of meaning that may be found in adverb-adjective combinations, I will venture beyond my corpus:

> *Cleverly incompatible* giver and taker
> ... Given their gross *incompatibilities*, it is possible to endorse their weird complacency. As Vita put it: "I do think we have managed things *cleverly*."
> (heading and end of text from a review in the *Times Literary Supplement*, June 26 (1992), p. 10)

At the outset, the heading seems opaque, but it neatly summarizes the point expressed in the final sentences of the text.

## 4. Collocations

Adverbs are no doubt the most heterogeneous of the traditional word classes. Syntactically, the patterns of co-occurrence are less marked than for the other classes of lexical words, as shown by a study of tag combinations in the LOB Corpus (see Johansson and Hofland 1989). Is this lower degree of syntactic dependence reflected in a lower degree of collocational patterning? It is certainly remarkable that adverbs are not listed as headwords in Benson *et al.*'s (1986) *Combinatory Dictionary of English*. But there is no doubt that adverbs enter into collocational patterns; see the detailed study by Bäcklund (1973) and the recent discussion of some aspects of adverb collocations in Altenberg (1991).

The original aim of this paper was to focus on collocational patterns, based on the LOB Corpus material (following up the work of Johansson and Hofland 1989). It quickly became apparent that the material was too small for a proper study of collocations, substantiating John Sinclair's claim that we need far larger corpora (see e.g. Sinclair 1982, 1991)[6]. Nevertheless, some remarks on collocational patterns may be in order.

Adverbs of degree range from highly 'grammaticized' intensifiers (Bolinger 1972:22), which collocate with a wide range of adjectives, to those which are severely restricted in use. These are some examples which occur repeatedly in the LOB Corpus (frequencies are given within parentheses; note also the examples given in Section 2):

> broadly comparable (2), comparatively new (3), comparatively small (4), deadly dull (3), deeply concerned (2), desperately worried (2), diametrically opposite (2), eminently respectable (3), equally good (5), equally important (4), exceptionally cheap (2), exceptionally high (2), extremely difficult (7), fairly accurate (3), fairly certain (3), fairly small (4), fairly wide (3), far better (9), far greater (7), fully aware (4), fully representative (2), heavily dependent (2), highly complex (3), highly important (3), highly significant (4), highly successful (2), immediately recognisable (2), increasingly important (3), just sufficient (2), long overdue (4), newly independent (5), particularly difficult (4), particularly important (9), perfectly happy (2), perfectly normal (2), perfectly true (2), plainly visible (2), pretty good (3), pretty obvious (4), pretty sure (4), purely formal (2), purely legal (2), purely physical (2), purely technical (2), quite alone (2), quite amusing (2), quite capable (3), quite clear (10), quite impossible (7), quite sure (14), quite true (4), quite unnecessary (3), rather disappointing (3), rather fine

(3), rather proud (3), rather surprising (2), readily available (2), really cold (2), really good (5), really important (3), reasonably careful (2), reasonably safe (2), relatively high (7), relatively low (5), relatively short (3), relatively simple (3), relatively small (7), remarkably good (2), remotely possible (2), seriously ill (2), slightly ridiculous (4), somewhat greater (2), statistically significant (4), sufficiently important (3), sufficiently large (5), sufficiently small (2), theoretically possible (3), thoroughly good (2), vastly superior (2), virtually impossible (3), vitally important (2), vitally necessary (2), well aware (9)

To be more meaningful, these figures should be related to the total frequency of the respective adverbs and adjectives in adverb-adjective combinations. We find then, for example, that *eminently* is only recorded in combinations with *respectable*, while *respectable* is preceded by *eminently* in three out of a total of six combinations. In other words, there appears to be a high degree of association between these words.

As numbers are so small, there is not much point in making exact calculations of collocational tendencies for individual word pairs. It may be more meaningful to see what observations can be made for types of words combining with a particular adverb or adjective, as Bäcklund (1973) does in his collocational study. For example, these are some adverbs that seem to occur with adjectives denoting positive and negative qualities, respectively:

brilliantly, enormously, immensely, reasonably, remarkably, supremely, wonderfully

bitterly, blatantly, dangerously, deadly, desperately, grossly, seriously, severely, singularly

Quirk *et al.* (1985:469) observe that *perfectly* tends to combine with positive adjectives and *utterly* with negative ones, though they point out that there are exceptions[7]. Such exceptions are found in the LOB Corpus material: *perfectly beastly/commonplace/ridiculous*, *utterly fearless/reliable*. This brings us to our next point.

## 5. Sense developments

The rise of grammaticalised intensifiers from words with a richer meaning has been frequently noted in the literature, e.g. by Stern (1965:393), who regards this as a case of adequation, one of his seven classes of sense change. Building on Stern's work, Warren (1992) describes the development as a case of implication, as shown in her analysis of *awfully* (p. 174):

Mother sense: 'to a degree that causes awe'
By implication: >'to an extremely high degree'> 'to a high degree'

Because of repeated use, the new meaning became conventionalised. The final stage is reached by adequation, which 'involves an adaptation of the meaning of a word so that it agrees with the language-user's perception of the actual characteristics of the referent' (Warren 1992:7). In other words, the original components of meaning are no longer attended to, and adequation sets in and makes sure that sense and reference agree.

Stern (1965:311) points out that many intensifiers 'have not yet become completely adequated, but retain more or less their primary meaning, and ..., for that reason, give more force and vividness to speech'. This applies to many adverbs which express attitude as well as degree (3.9), and it may be reflected in collocational patterns, as in the case of *perfectly* (cf. Section 5).

By referring to implication and adequation, we can thus account for changes in meaning of degree adverbs. At the same time, we may gain more insight into the complexity of meaning and collocational patterns of adverbial modifiers in general.

## 6. Concluding remarks

The seeming simplicity of adverb-adjective combinations disguises complexities of meaning which are normally unnoticed. How do speakers get their messages across? The language user has available a vast store of words and expressions, with their meanings and combinatory potential, based on previous exposure to language. These form the basis of language production and interpretation. Sometimes the interpretation is highly restrictive, e.g. with collocations like *eminently respectable* or *blissfully ignorant*. At other times it has to be computed with reference to productive rules and the surrounding text, as in our examples with *safely dead* (3.9) and *cleverly incompatible* (3.10).

In this paper I have merely scratched the surface of an area which merits far more exploration. The LOB Corpus material is rich enough to show the wide variety of adverb-adjective combinations, but insufficiently large for a proper collocational study. For this we need the vast corpora and the analysis tools currently under development, largely inspired by John Sinclair's work.

## Notes

1. I am grateful to Knut Hofland, Norwegian Computing Centre for the Humanities, for assistance in retrieving material for the study. For comments on an earlier version of this paper I am indebted to my colleagues Hilde Hasselgård and Per Lysvåg, Department of British and American Studies, University of Oslo.
2. More precisely, the material retrieved was: all expressions containing a word plus a tag beginning with RB followed by a word plus a tag beginning with JJ. Note that a number of common degree adverbs (including very) have a special tag and were therefore excluded from the material. See Johansson et al. (1986:68ff.).
3. References to the corpus contain a letter code for the text category, followed by the number of the text in the category, and the line number in the text. See further Johansson *et al.* (1986:1, 5).
4. Cf. Bolinger (1972:23): 'Any list (of intensifiers) has to be viewed as a sampling rather than as a catalog, not because the set of intensifiers is too big to do more than sample, but because it is too open-ended.'
5. As regards the relationship between adverbs as modifiers and as sentence adverbs, see especially Swan (1982, 1988).
6. Another way of coping with the scarcity of material for individual words is to supplement the corpus with material elicited from native informants, as Bäcklund (1973) does in his study of adverb collocations.
7. Bäcklund (1973:226) observes that the 'heads in the range of *perfectly* are all (except *horrible*, *indifferent* and *unthinkable*) in themselves positive or commendatory in their denotations, but the sentences are to a large extent sarcastic in nature.' With respect to *utterly*,

he notes that the informants agreed that it has 'an essentially derogatory connotation' (p. 214), but his material does include examples of *utterly* with a neutral or positive import.

## References


Allerton, D J. 1987. English intensifiers and their idiosyncrasies. In Ross Steele and Terry Threadgold (eds) *Language Topics: Essays in Honour of Michael Halliday*. Vol.2. Amsterdam/Philadelphia: John Benjamins Publishing Co. pp 15-31

Altenberg, Bengt. 1991. Amplifier collocations in spoken English. In Stig Johansson and Anna-Brita Stenström (eds.), *English Computer Corpora: Selected Papers and Research Guide*. Berlin: Mouton de Gruyter. 127-147.

Bäcklund, Ulf. 1973. *The Collocation of Adverbs of Degree in English*. Studia Anglistica Upsaliensia 13. Stockholm: Almqvist & Wiksell.

Benson, Morton, Evelyn Benson, and Robert Ilson. 1986. *The BBI Combinatory Dictionary of English: A Guide to Word Combinations*. Amsterdam: Benjamins.

Bolinger, Dwight. 1972. *Degree Words*. The Hague: Mouton.

Borst, Eugen. 1902. *Die Gradadverbien im Englischen*. Heidelberg: Carl Winters Universitätsbuchhandlung.

Johansson, Stig, Eric Atwell, Roger Garside, and Geoffrey Leech. 1986. *The Tagged LOB Corpus: Users' Manual*. Bergen: Norwegian Computing Centre for the Humanities.

Johansson, Stig and Knut Hofland. 1989. *Frequency Analysis of English Vocabulary and Grammar*. Vol. 2: *Tag Combinations and Word Combinations*. Oxford: Clarendon Press.

Johansson, Stig and Per Lysvåg. 1987. *Understanding English Grammar*. Vol. 2. Oslo: Universitetsforlaget.

Jonson, Gördis. 1967. Adverbs of degree in the *Observer*. *Moderna Språk* 61, 337-353.

Kirchner, Gustav. 1955. *Gradadverbien, Restriktiva und Verwandtes im heutigen Englisch*. Halle: VEB M. Niemeyer Verlag.

Poldauf, Ivan. 1959. Further comments on Gustav Kirchner's *Gradverbien*. *Philologica Pragensia* 2, 1-6.

Quirk, Randolph, Sidney Greenbaum, Geoffrey Leech, and Jan Svartvik. 1985. *A Comprehensive Grammar of the English Language*. London: Longman.

Sinclair, John. 1982. Reflections on computer corpora in English Language research. In Stig Johansson (ed.), *Computer Corpora in English Language Research*. Bergen: Norwegian Computing Centre for the Humanities.

Sinclair, John. 1991. *Corpus, Concordance, Collocation*. Oxford: Oxford University Press.

Sinclair, John *et al.* (eds.). 1990. *Collins COBUILD English Grammar*. London: HarperCollins Publishers.

Spitzbardt, Harry. 1954. *Die modernen Gradadverbien. Ein Beitrag zum englischen Sprachgebrauch des 20. Jahrhunderts*. Diss. Jena.

Spitzbardt, Harry. 1965. English adverbs of degree and their semantic fields, *Philologica Pragensia* 8, 349-359

Stern, Gustaf. 1965. *Meaning and Change of Meaning with Special Reference to the English Language*. Bloomington: Indiana University Press. (First published in Sweden in 1931.)

Swan, Toril. 1982. A note on the scope(s) of *sadly*. *Studia Linguistica* 36, 131-140.

Swan, Toril. 1988. *Sentence Adverbials in English: A Synchronic and Diachronic Investigation*. Oslo: Novus.

Warren, Beatrice. 1992. *Sense Developments*. Stockholm Studies in English 80. Stockholm: Almqvist & Wiksell International.

# WHERE TO *BEGIN* OR *START*?: ASPECTUAL VERBS IN DICTIONARIES

*Gerhard Leitner*
Freie Universität Berlin

## 1. The grammar-lexis interface

The interface between grammar and lexis and the consequences that arise out of relating them in reference materials have been central to Sinclair's work. To quote from a paper at the international conference on grammar-writing in Berlin in 1986:

"In the explicit theoretical statements of linguists, grammatical and lexis patterns vary independently of each other. In most grammars it is an assumption that is obviously taken for granted. For example it is rare for a grammar to note that a certain structure is only appropriate for a particular sense of a word. The same goes for morphology. In contrast, grammars attribute independent meaning to syntactic arrangements.

Equally, it is rare for a dictionary to note the common syntactic patterns of a word in a particular sense. Pedagogical dictionaries are increasingly seeing this as essential information for learners, but it is added in the form of afterthoughts such as usage notes. The message of the conventional dictionary entry is that most of the words in daily use have several meanings, and any occurrence of the word could signal any one of the meanings.

The decoupling of lexis and syntax leads to the creation of a rubbish dump that is called idiom, phraseology, collocation and the like." (Sinclair 1986:59f)

*Looking up* (1987) makes these points even more strongly: "In nearly every case, a structural pattern seemed to be associated with a sense." (Sinclair 1987b:109) or "(T)here was in practice no clear distinction between grammar and lexis, and grammatical rules merged with restrictions in particular instances, and those restrictions ranged from the obviously grammatical to the obviously lexical." (1987:110). Meaning, collocation, phraseology, and syntax merge as one studies the behaviour of (particular) words and, vice versa, when one tries to define the classes of words that are able to function in given grammatical patterns. The *Collins COBUILD English Language Dictionary* (CCELD 1987) and the *Collins COBUILD English Grammar* (CCEG 1990) are clear consequences of that insight, the former by introducing the "extra column" with detailed grammatical information, which is also used as a criterion for distinguishing senses, the latter with lists of (classes of) lexical items that exemplify structural patterns. CCELD could be called a dictionary of (grammatical) patterns, CCEG a grammar of (lexical) classes.

In this paper I will address two problems: firstly, how the grammatical part in dictionaries can be improved upon so that the link-up with grammars is facilitated; secondly, how close CCELD (and other dictionaries) come to represent "real

English", and what else needs to be said, and where. The argument will rest on the assumption that lexical entries in dictionaries rely on the following constitutive elements and their interplay:

A) grammatical codes to describe sentence syntax,
B) a vocabulary for the semantics of sentence constituents,
C) for verbs, a vocabulary for the role of participants associated with, and process(es) expressed by, the verbs,
D) examples to illustrate or further define sense(s),
E) the number and order of sense distinctions.

I will look at CCELD, the *Longman Dictionary of Contemporary English* (LDOCE 1987), and the *Oxford Advanced Learner's Dictionary* (OALD 1989). To be practical, I will focus on the aspectual meanings of *begin* and *start*. These verbs are paradigmatic cases for scores of similar ones, such as *commence*, *finish*, *end*, but also *want*, *force*, *try*, etc.

## 2. Grammar in contemporary advanced learner's dictionaries

### *2.1. Some general issues*

Since the publication of *Hornby's Advanced Learner's Dictionary* in the late 1940s, pedagogic dictionaries have included explicit information on whether such syntactic functions as objects, complements, and adverbials are obligatory or optional with some verb. In doing that, they unanimously focus on *surface structure* patterns. Explicit information on the way(s) these constituents are realized, *viz.* as noun phrases, prepositional phrases, and/or clauses, is more recent and dealt with in different ways. But it is in the coverage of more detailed syntactic matters that contemporary dictionaries differ crucially from one another. This holds, for instance, for the marking of sentence processes or relationships such as active/passive, the ergative correspondence (CCELD, CCEG), the coverage of restrictions on, or preferences for, a particular pattern (e.g. active or passive), the mention of prepositions with adverbials, and the use of frequency data.

There are other areas of divergence. Grammatical codes allow the (informed) user to generalize from the 'specific' into the 'grammatical'. Those generalizations rest on clear and explicit definitions. Dictionaries differ in the degree to which they allow such inferences. Furthermore, there is a trade-off that is often overlooked (Hausmann 1957) between the above-mentioned elements A) to D) so that the 'correct' or appropriate use of a lexical item can, and often is, to be inferred from several different sources of information.[1] In LDOCE, for instance, it is the function of examples, category D) in section 1, to illustrate the grammatical potential of an item.

In other words, the coverage of grammar in dictionaries is not uni- but multidimensional and 'invited' inferences to be drawn by the user are based on a complex of decisions. The user-friendliness and efficiency of dictionaries depends on the knowledge of the intricate web of relations.

### *2.2. The grammar of 'begin' and 'start'*

The grammar and meaning of *begin* and *start* is well understood in general terms (CCEG 1990, Quirk *et al.* 1985, Palmer 1987, Gramley 1988). Both verbs can be used transitively and intransitively, and they allow non-finite complement clauses.

*Start* can be used causatively, *viz.* "He started the race", a use also mentioned in Dixon with *begin*, as in "The master began the boys running as they passed the copse" (1991:177). From a semantic point of view, they have been called aspectual verbs since they refer to the initial stages of events, etc. They have other meanings as well. Thus *start* has a movement sense, as in "He started up quickly", *begin* functions as a quasi-speech act verb of saying, as in "I'm terribly sorry, he began". They are by no means total synonyms. Usage and frequency differences have been observed with regard to the choice of complement clause type (Legler 1975). Theoretical issues remain on whether they are to be analyzed like *seem* with subject or like *try* with object clause embedding or as both, as is argued by Perlutter (1979). Dixon (1991) raises the question whether most intransitive and transitive uses should not rather be analysed as implying semantically triggered subordinate verb phrase and/or object deletion (in conjunction with object raising etc.) to account for the semantic predictability.

The question now is how these verbs are treated in the dictionaries under consideration. As implied above, there is no scope for a comprehensive comparison of the coverage of the grammar of these verbs; the following tables and diagram merely attempt to capture the major aspects of *begin* in the three dictionaries (diagram 1), of *begin* and *start* in CCELD (table 1) and of *start* in CCELD and LDOCE (table 2).

| BEGIN | | | START | |
|---|---|---|---|---|
| no. of sense | syntactic code | paraphrase of meaning | syntactic code | no. of sense |
| 1 | V-to/ing | you _ to do/feel (or doing/feeling) sth | V-to/ing V (O) (corr.) | 1 |
| 2 | V–ERG (corr.) | sth _s or you _ it (# EVENT) | V–ERG | 2 |
| 3 | V(O) A *by-ing, with* NP | you _ (# ACTIVITY, PROCESS) *with, by doing* sth | V (O) A *by-ing, with NP* | 6 |
| 6 | V(O)A _ *as* NP | sth _s *as* another | ./. | POS |
| 7 | V(O)A _*as* NP | so _ed career *as* NP | V(–PHRASAL) (O)A *as* etc. (corr.) | 7 |
| POS | ./. | you _ sth, such as (new business etc.) | V(–PHRASAL) O A (corr.) | 8 |
| 9 | V(A) (corr.) | (# SPATIAL) N _ AdvP | V(A) (corr.) | 13 |
| 10 | V (O) (A-*with* NP) | sth printed/ written _ | ./. | POS |
| IMP | ./. | you _ an engine motor, car etc. | V(–PHRASAL)–ERG A (corr.) | 9 |
| IMP | ./. | you _ AdvP (# MOTION) | V A | 10 |
| IMP | ./. | you _ a race (causative) (# EVENT) | V O | 12 |
| 5 | V (–QUOTE) | you _ (saying) | ./. | POS |

Table 1: Comparison of *begin* and *start* in CCELD

Abbreviations in table: _ = *start* or *begin*; POS = possible sense not included in CCELD; IMP = impossible sense; so = someone; sth = something; a combination of brackets (..)(..) means that one or both constituents must be present; # and corr. indicate further changes made by the author of this paper.

Table 2: Comparision of *start* in CCELD and LDOCE

| CCELD | | | LDOCE | | |
|---|---|---|---|---|---|
| no. | code | paraphrase | paraphrase | code | no. |
| 1 | V (0)<br>V-to/-ing | you _ to do/feel sth | (cause to) go into state of (movement, operation, activity) | I, T ( OFF, *with, by-ing*) (corr.) | 1 |
| 2 | V-ERG | sth –s or you _ it | | (T)+*to* – v/v–*ing* | |
| 6 | V(O) A *by – ing*<br>*with* NP<br>*missing entirely | you –sth (#: ACTIVITY, PROCESS) | | *(T)+obj. + v-*ing* | |
| 2 | V–ERG | sth_s or you_it | (cause to) come into existence | I, T (UP) | 2 |
| 8 | V(–PHRASAL) OA (corr.) | you – sth, such as new business etc. | (# = go into state of being) | | |
| 9 | V(–PHRASAL) ERG + A (corr) | you _an engine, motor, car etc. | (cause to) begin operation (# = go into state of) | I (UP), T | 3 |
| 1 | V (only this)<br><br>(V+*on* is a separate entry) | you _ | begin doing a job or piece of work (# = go into state of v-*ing*) | I (IN, *on*) | 4 |
| 10 | V+A (partial mapping | you _ somewhere | begin a journey (# = go into a state of activity) | I (OFF, OUT, *for*) | 5 |
| (13?) | V+A | where a region etc.<br>_s | go from a particular point (# = go + adv.) | I + adv/prep. esp. *at, from* | 6 |
| | no mapping | | begin using (# 0 go into state of v-*ing*) | T | 7 |
| (7) | V(-PHRASAL) OA (corr.) (partial map) | so _ed career etc.<br>*as* | begin one's life, course of action | L (OFF, OUT) | 8 |
| | no mapping | | take part in a match or competition from beginning | I | 9 |
| 12 | V+O | you _ a race | | no mapping | |

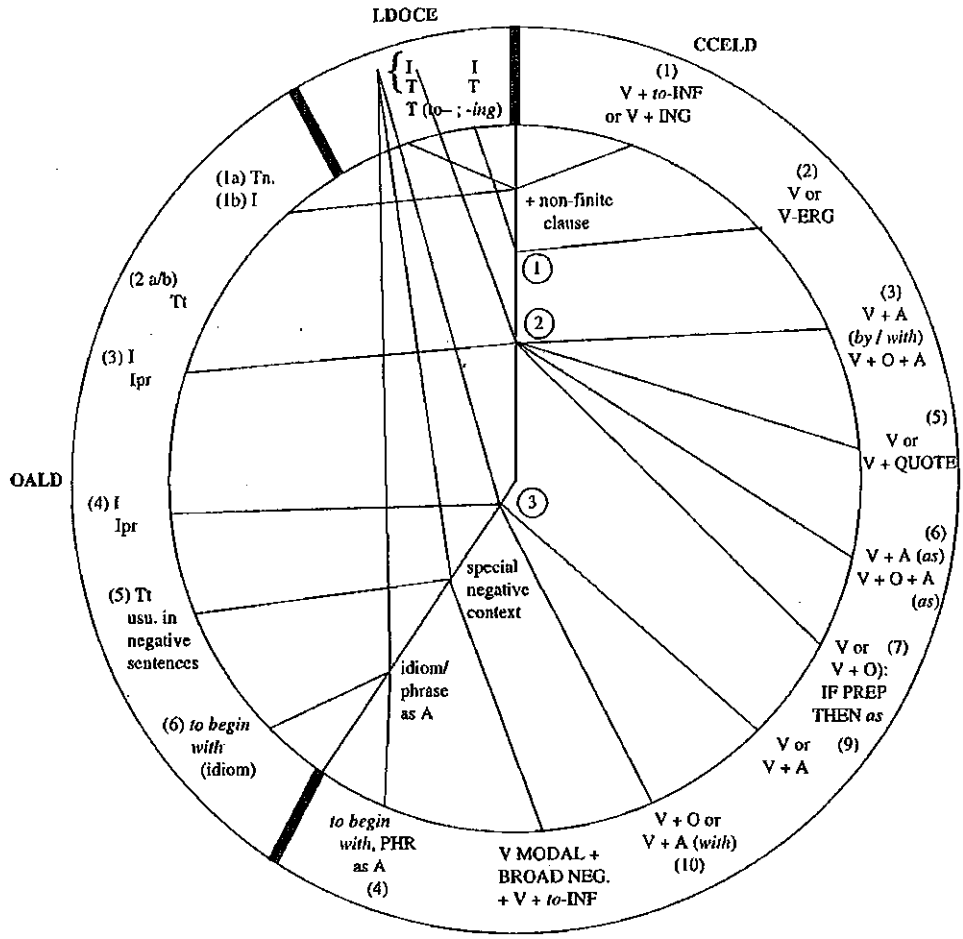



Diagram 1: Grammatical treatment of *begin* in the three dictionaries

KEY TO DIAGRAM

**CCELD** (Collins Cobuild English Language Dictionary)

| | | |
|---|---|---|
| V | : | Verb |
| INF | : | Infinitive |
| ING | : | Verb ending with *-ing* |
| V-ERG | : | Ergative verb |
| A | : | Adjunct |
| O | : | Object |
| QUOTE | : | Quoted speech |
| PREP | : | Preposition |
| PHR | : | Phrase or expression |

**LDOCE** (Longman Dictionary of Contemporary English)

| | | |
|---|---|---|
| I | : | Intransitive |
| T | : | Transitive Verb |

**OALD** (Oxford Advanced Learner's Dictionary)

| | | |
|---|---|---|
| Tn | : | Transitive verb taking direct object (noun phrase or pronoun) |
| I | : | Intransitive verb |
| Tt | : | Transitive verb taking *to*-infinitive clause as direct object |
| Ipr | : | Tntransitive verb followed by prepositional phrase |

Diagram 1 shows that OALD makes use of an elaborate set of codes for (phrasal and clausal) complementation, adverbial and other necessary or optional syntactic functions, e.g. 'Tt, Tg, Ip, Ipr, Cn'g'. They are well explained and amply illustrated in the appendix. Frequently, a specific, but *broadly defined*, sense is associated with several codes or, to put it another way, it can be realized by a range of grammatical patterns. LDOCE is similar in postulating broad sense distinctions with a wide range of grammatical patterns. But it subsumes the entire grammatical potential just under two labels, 'T' and 'I', for transitive and intransitive use. The wider potential grammatical patterns must be inferred from examples. Additional, but subordinate, codes inside sense 1 do, at least, indicate the possibility of clausal complementation separately, *viz.* (*+to v*) and (*+v-ing*). In contrast to both dictionaries, CCELD differentiates senses on the basis, at least in part, of particularities in syntactic potential. It thus has the greatest number of senses, which often seem to be quite closely related.

Diagram 1 also highlights other areas where the three dictionaries agree and, more importantly, where they differ. Notice the pervasive differences between CCELD on the one hand and LDOCE and OALD on the other in the coverage of (optional) adverbial constituents (on *start* cf. tables 1 and 2).

There are several codes that are specific to some dictionary, a fact that reflects either the inclusion (or exclusion) of different data or divergent views of the structure of English. The first possibility is exemplified by LDOCE's use of the concept of link verb (L) to account for sentences like (1a):

(1) (a) He started poor/a poor office boy
(b) He started poor/*as* a poor office boy.

None of the other dictionaries mentions such data. LDOCE and OALD omit sentences like (1b), with an adverbial phrase introduced by *as*, which is mentioned in CCELD.

The second possibility is illustrated in the way the dictionaries handle sentence pairs like (2) and (3):

(2) The rumour started immediately
(3) They started the rumour immediately

LDOCE handles this pair in terms of mood and causativity. It describes sense 1 of *start* as "to (cause to) go into a state of (movement, operation, or activity)" (causativity). OALD on the other hand handles such pairs in terms of transitivity. Commenting on an equivalent pair of uses of *begin*, it describes *begin's* sense 1(a) as "set (sth) in motion" and 1(b)"be set in motion" (mood), which is illustrated by such examples as "begin work" and "building began last year". CCELD uses the concept of ergativity (code 'V-ERG') to describe the relation, which implies mere agentivity. Ergativity seems a reasonable way of relating such sentence pairs and can lead to a deeper level of syntactic generalization, provided the possible range of constituents in subject or object position can be delimited.

It was said earlier that grammatical generalization rests on a clear interpretation of the codes and their interaction. Information in this area is conflicting in and between the dictionaries. In its gloss on Tt/g OALD, for instance, correctly implies that a *to*-infinitive and an *ing*-participial clause reflect the transitive character of the verb[2]. LDOCE is confusing in this respect. The grammatical information for the *entire* entry of *begin* is (I;T), meaning that *begin*, like *start* (cf table 2), can be

both transitive and intransitive. The subordinate code (*+to v*) inside sense 1 of *begin* can therefore be interpreted ambiguously and no clarification emerges from the dictionary guide (F41) either. The treatment of *start* is similarly deficient. CCELD's coverage suffers from the fact that there are no special entries for the codes 'V+*to*-INF', or 'V+*ING*', and there are inexplicable differences between *begin* and *start* (compare senses 1 each).

To turn to the semantic description of sentence constituents, ( B) in section 1), and the process implied by, or suggested with, particular participants in subject and/or post-verb position, ( C) in section 1), all dictionaries characterize subjects, object and adverbs/adjuncts in terms of semantic descriptors, such as animate and/or human (*someone, you*), or inanimate (*something*), or by using more specific vocabulary, such as *region, career, course of action, engine*. But there is a clear difference in degrees of specification between the subject and post-verb constituents, which of course reflects the facts of the semantics and grammar of aspectual verbs.

Interestingly, sense ordering seems to reflect somewhat degrees of semantic specificity of participants, and that quite independently of the differences in approach between CCELD on the one hand and LDOCE and OALD on the other. This observation is, of course, a consequence of the distinction between broadly defined senses and more specific, but related, sense distinctions. As far as subjects are concerned, there is only a weak tendency of senses going from general to more specific descriptors. To illustrate this from table 2, CCELD's senses 1,2,3,6 to 10,13, and 16 (the verbal senses) imply an animate/human subject "you" (1,3,6-10,12), "someone" (7), or "something"(2). Sense (13) has inanimate subjects and more specifically spatial nouns, such as "region, place, type of countryside, etc.". (16) has "a man and woman", i.e. a very specific choice of subject indeed. In LDOCE subjects are hardly defined at all by descriptors and only clarified through the examples. Thus, sense (1) combines examples with animate-human and inanimate subjects. Sense (2) has only animate-human subjects, etc.

Objects, complements, and (in CCELD's case) adverbials reflect a broader spectrum of semantic differences, and a more pronounced tendency to move from more general to more specific descriptors. Thus, CCELD has "something" in (1),"something/it" in (2), an implied activity or process in (6), etc. OALD starts the same way but then mentions states of mind in b.

To turn to examples D) in section 1, the difference in attitude towards data between CCELD and CCEG on the one hand and LDOCE (and, maybe, OALD) on the other is well-known (Sinclair 1987a). For Sinclair data "support the explanations and they illustrate usage. They provide a reliable guide for speaking and writing in the English of today." (CCELD 1987:xv). And he goes on to say that,"in contrast, invented examples are really part of the explanation. They have no independent authority or reason for their existence." To put it slightly differently, examples in CCELD are to help the user either to assign a particular (occurring) example to a sense or to lead him to construct correct 'real' English sentences or utterances.

Examples in CCELD are rarely edited. More importantly, since they do, and have to, stand on their own, they make one point, not several, and they are not commented on or paraphrased. In OALD and LDOCE, examples often further define a sense, and are then glossed or paraphrased, or they even serve multiple

purposes. Here are some illustrations. "[*S*]*tart* (ie begin using) *a new tin of paint*" (OALD, sense 2, glossing), and "*We'll begin by dancing/with a story/at the beginning*" (LDOCE, sense 1, multiple purpose example). There is another difference between CCELD and LDOCE and OALD, namely the richness of data in CCELD (and CCEG, for that matter), cf. table 3:[3]

| | CCELD | LDOCE | OALD |
|---|---|---|---|
| *begin* | 28 | 9 | 23 |
| *start* | 27 | 28 | 17 |
| total: | 55 | 37 | 40 |

Table 3: Number of examples in three dictionaries

While LDOCE has comparable clausal codes, *viz.* A), and (partially) overlapping paraphrases, *viz.* C), in section 1 for *begin* and *start*, differences in the semantic roles or selection restrictions associated with the participants are based on the semantic analysis of the nouns in question. Thus, senses (1) and (2) of *start* have an optional causative reading; they differ in that (1) implies a transition "into a state of (movement, operation, or activity)" and (2) a coming "into existence". In other words, according to (1), one goes into a state of movement, referred to as a 'journey' or an activity called a 'meeting' (examples of sense 1), while (2) leads to something new ('object of result') that can be referred to as 'trouble', 'rumour' or 'swimming club' (examples of sense 2). Sense (3) can be paraphrased as a transition into an "operation", such as an engine starting, or starting an engine. As one goes down the list of senses, selection restrictions or readings become more and more specific. (5), an intransitive use of *start*, has the paraphrase "begin a journey", (6), an intransitive use with an obligatory adverbial (of place or time), "go from a particular point", (7), which is transitive, has "begin using", (8) "begin one's life, a course of action, etc.". OALD's treatment is comparable but confines its analysis to just five senses, as against LDOCE's eight.

CCELD is somewhat different in these respects. While the nature of the participants involved does play a role, as in the causative sense 12 of *start*, *viz.* "start a race", sense divisions are primarily associated with differences in syntactic patterns, which are *not* limited to obligatory constituents. Thus (1) has "V, V+O, V+*to*-INF/*-ing*", which express the meaning of "if you **start** to do or feel something, ...". (2) has the code "V-ERG" to express the relationship between things one can start and that can start by themselves. The further paraphrase makes it clear that nouns must designate an event, an activity, or operation. Moreover, senses are distinguished on the basis of associated adverbials or adjuncts. Thus, senses (6) and (7) involve adverbial NPs with *with* or *as*, or adverbial clauses with *by -ing*, witness

(4) He started his scientific career by playing around with pins and threads
(5) I started with a joke
(6) She started as a secretary

Adverbials are treated as optional constituents in LDOCE and OALD and hence subsumed under relevant senses of *start* as a transitive or intransitive verb. Note that CCELD does not claim that they are obligatory in any way, but that they frequently pattern with minor sense distinctions in the corpus. Hence, a number of sense distinctions are based, more or less, entirely on them.

## 3. Some problems and more data for pedagogical dictionaries

The survey of coverage of the grammar of *begin* and *start* raises a number of problems. Only two will be dealt with here since they have a broad impact on lexicographic practice. They have to do with differences in linguistic approach on the one hand and a very general linguistic problem on the other. I will also mention some facts that derive from the LOB corpus data and bear upon a more complete description of the use of these verbs.

The first problem concerns the analysis of sentences, such as (2) and (3) above. They are treated as manifestations of ergativity in CCELD and of transitivity in LDOCE and of transitivity, in conjunction with mood, in OALD. No analysis is completely convincing, but for different reasons. CCELD is unable to demonstrate clearly syntactic differences between other transitive and intransitive pairs or, at least, the classes of nouns that can enter the ergative relationship. It seems to imply an agentive reading for (3) that differs from a 'true' causative one. That is postulated for *start*'s sense 12, cf. (4):

(4) He started the race

That sentence must be assumed to mean "caused the race to start" or "caused the race to start to take place". And as the paraphrases show, this implies non-identity of matrix subject and (embedded) subject. The matrix object in (4), *the race*, results from an embedded subject position after the verb phrase has been deleted. Such an analysis leads on to the second problem below.

LDOCE and OALD treat (2) and (3) as manifestations of verb processes, i.e. as causativity, agentivity, processuality, and mood. Again, no clear picture emerges. Thus, LDOCE'S paraphrase of sense 1 of *begin* is "to do or be the first part of", which implies an agentive/processual relationship and OALD's senses 1a and 1b are "set something in motion" and "be set in motion". LDOCE's sense 2 of *begin* also seems to rely on something similar to CCELD's ergativity when the paraphrase reads as "to (cause to) come into existence". It is illustrated by

(5) Mary began a club for bird-watchers

and is paired with *start*'s paraphrase of sense 1, i.e. "to (cause to) go into a state of". Agentivity, processuality and causativity are referred to elsewhere in LDOCE but no clear picture emerges of the syntactic and/or semantic differences, nor of the nature of the semantic differences.

Another more general problem for all dictionaries is that they fail to capture relationships that are, or at any rate can or should be, better explained in terms of ellipsis. Sentence (4), on one analysis of the causative reading, already illustrates the problem in general. But the problem is of greater import as is shown by Dixon (1991). To give a few examples of what he considers to be relevant cases (deletable constituents are in brackets):

(6) The choir started (singing) 'Messiah' at two o'clock
(7) Mary continued (writing) her book after a short holiday
(8) Tommy had finished (shelling) the peas
(9) I've completed (grading) these assignments
(10) He began (cooking) the supper
(11) She began (knitting) a sweater
(12) My uncle began (telling) another joke

(13) a) The chef started cooking the dinner at four o'clock
b) The chef started (cooking) the dinner at four o'clock
c) The chef started (cooking) (the dinner) at four o'clock
(14) a) John's jealousy began when he saw Mary out with Tom
b) John began to be jealous when he saw Mary out with Tom

Sentences (6) to (12) illustrate verb phrase deletion. (13) has both verb phrase and (embedded) object deletion which results in either transitive surface structure patterns or in intransitive ones. Like (13), (14) shows that intransitive patterns can be the result of verb phrase deletion with so-called activity nouns. Dixon argues convincingly that many patterns may be understood in terms of consitituent deletion and processes of object raising etc. They are, he argues, triggered by semantic principles. He disagrees with Sinclair's use of the notion of ergativity, following Halliday, and argues that pairs like (2) and (3) or (16) should be understood in terms of raising to subject:

(15) a) John read that novel
b) That novel reads well
(16) a) John began (reading) that novel
b) That novel begins well
c) (*That novel begins reading well)

(16a) with *reading* deleted and (16b) are only possible in cases where (15a/b) are possible. The role of Dixon's semantic approach to (English) grammar deserves to be studied more fully since it also presupposes semantically-based classes of nouns, adjectives and verbs. But its application to dictionary practices may require a radical step away from focussing exclusively on surface structure unless a way is found to cross-refer the (willing) user to a complementary grammar.

To close with some facts on frequency patterns that are covered neither in dictionaries nor in grammars. There are two assumptions that are arguable. The one is that the *to*-infinitive and *ing*-complement clauses are synonymous, the other that they may occur with equal frequency with both verbs. While the former assumption is by no means uncontroversial (Dixon 1991), I will limit myself to the latter assumption, which incidentally, is not made in LDOCE.

The LOB corpus clearly shows that *begin* and *start* behave quite differently syntactically, cf. tables 4 to 7:

| | *begin* | | *start* | |
|---|---|---|---|---|
| verb base (VB) | 86 | 17.59% | 116 | 34.63% |
| 3rd p. sg. (VBZ) | 30 | 6.13% | 21 | 6.27% |
| *ing*-form (VBG) | 50 | 10.22% | 38 | 11.34% |
| past tense (VBD) | 271 | 55.42% | 116 | 34.63% |
| past part. (VBN) | 52 | 10.63% | 44 | 13.13% |
| TOTAL | 489 | | 335 | |

Table 4: Frequency of verb forms (tokens) (grammatical code of LOB in brackets)

It is evident from this table that the verb forms do not occur with an evenly distributed frequency. Table 7 will reveal that both verbs are primarily past time verbs, although *start* is not as extreme in this respect as *begin*.

To turn to the frequency of post-verb syntactic patterns:

| | | *begin* | | *start* | |
|---|---|---|---|---|---|
| I. | + *to*-compl. Cl. | 261 | 53.4% | 37 | 10.8% |
| II. | + *ing*-compl. Cl. | 24 | 4.9% | 53 | 15.5% |
| III.1. | + *by* V-*ing* Cl. | 11 | 2.2% | 5 | 1.5% |
| III.2. | + *with*+NP | 23 | 4.7% | 23 | 6.7% |
| III.3. | + *as*+NP | 3 | 0.6% | 4 | 1.2% |
| IV. | intrans. use | 109 | 22.3% | 116 | 33.9% |
| V. | trans. use | 45 | 9.2% | 104 | 30.4% |
| VI. | *to*+V+*with*-Adjunct | 13 | 2.7% | 0 | 0.0% |
| TOTAL | | 489 | | 342 | |

Table 5: *begin* and *start* compared according to major syntactic patterns (distinguished in CCELD)

While *begin* mainly occurs in complex clause constructions with complement clauses (catg. I and II together amount to 58.3%), *start* shows up predominantly in monoclausal patterns (catg. IV and V amount to 64.3%). I will look into the grammatical analysis of categories III. 1-3 below but let me say here that they belong to the monoclausal pattern. They account for 7.5% of *begin* and for 9.4% for *start*. Taking this into account, 39.0% of the tokens of *begin* occur in a monoclausal constructions (catg. III.-V.) and 73.7% of *start*.

As for monoclausal patterns, *start* is used about equally frequently as a transitive and an intransitive verb, but *begin* is mainly used intransitively. As for complement clause types, *begin* prefers the *to*-infinitive clause, while *start* is more balanced, with the *ing*-clause being more frequent even than the infinitive.

The frequency of verb form tokens with complex, complement clause constructions is equally telling:

| | *to*-compl. | | *ing*-compl. | |
|---|---|---|---|---|
| | *start* | *begin* | *start* | *begin* |
| base form | 7 | 31 | 23 | 1 |
| 3rd p. sg. | 3 | 11 | 2 | 0 |
| *ing*-form | 3 | 38 | 1 | 0 |
| *ed*VBD-form | 17 | 161 | 25 | 23 |
| *ed*VBN-form | 7 | 20 | 2 | 0 |
| SUBTOTAL | 37 | 261 | 53 | 24 |

Table 6: Complement clauses with verb forms

There are a couple of features in which the two verbs coincide more or less. Firstly, passivization is extremely infrequent. There are only 13 cases with *start* and 7 with *begin*. Secondly, they are alike as regards time orientation. While the LOB corpus has unambiguous codes for 3rd person present and past tense, the base form (encoded VB) and the *ing*-form (encoded VBG) may reflect a variety of uses, such as imperatives, non-finite forms (after *did*, modal verbs, in complement clauses etc.) The details can be found in table 7:

| | *begin* | | *start* | |
|---|---|---|---|---|
| pres/fut. time reference: | 101 | 21.13% | 103 | 30.76% |
| past time reference: | 348 | 72.80% | 168 | 50.14% |
| non-fin., imper.: | 29 | 6.07% | 64 | 19.10% |
| TOTAL | 478 | | 335 | |

Table 7: Time orientation with *begin* and *start*

*Begin* and *start* are primarily verbs to refer to past situations but *begin* is much more extreme in this respect than *start*. On a smaller scale both are used to refer to future situations, i.e. in cases where the verb is part of a temporal clause, in a modal verb phrase, or (not counted here specifically) in a non-finite clause or an imperative. Only a very small portion actually refer to a straightforward present time. In these respect both verbs differ markedly from those studied by Kjellmer (1992).

## 4. Conclusion

Based on Sinclair's tenet that there is no difference in kind but only one of degree between lexis and grammar which can be captured in terms of the delicacy of the approach, I have looked at the way the grammar of the two verbs *begin* and *start* is covered in current pedagogical dictionaries. The following conclusions have been reached.

I have shown, firstly, that there is some trade-off between the constitutive elements of lexical entries. Thus, as Sinclair has argued (1986: 59), LDOCE and OALD frequently use examples to explain further the syntactical potential of the verbs. They are a necessary part of the syntactic description whereas CCELD's approach gives them an independent status. Connected with this is the fact that the semantic (or encyclopaedic) information on participants and processes amplifies on what could be called selection restrictions that border on syntax. Thus, while the relation between sentence pairs like (3) and (4) above is seen as one instance of transitivity in LDOCE and OALD, it is analysed independently in terms of ergativity in CCELD. Unfortunately, it is not made clear how that notion can be generalized and how such pairs differ from other sense entries.

The second point was that CCELD is strongest in terms of the delicacy used in grammatical description. Apart from the notion of ergativity, this holds in particular for the coverage of (optional) adverbial constituents.

Thirdly, I have argued that there remain several problems that are not covered or are hard to cover in dictionaries. This holds for a more precise delimitation of the classes that, for instance, enter into the ergative subject/object relationship and, secondly, for the deletion of constituents (ellipsis). Ellipsis yields similar syntactic surface structures where there actually are syntactic differences on a deeper level. As dictionaries focus on surface structures only, they can indicate these facts only in an artificial way (e.g. bracketing in examples), if at all. While the defect is a general one, it comes out most prominently in CCELD's coverage of optional adverbials which leads to the postulation of sense distinctions that go far beyond what is necessary for an adequate analysis. And yet the fact that those adverbials are noticed as a reflection of collocation makes it clear that the other dictionaries are missing a number of relevant points.

Fourthly, without going into the complex issue of a semantic analysis of corpus data, such data demonstrate convincingly crucial frequency differences which are overlooked in the dictionaries (though LDOCE does mention an *ing*-clause preference with *start*). The fact that these observations are not reflected in CCELD and CCEG does not seem to be accidental. The usage patterns shown in tables 5 to 8 do not reflect Sinclair's assumption about the probability of the co-occurrence of patterns and senses. As these patterns do not appear to correlate with semantic

distinctions, they find no place in that dictionary nor in the grammar. But where should they be treated then?

The four issues all point in one direction: the need for a more adequate descriptive linguistic basis and the design of dictionaries and grammars that interconnect. I believe that Dixon's semantically-based approach to English grammar (1991) indicates the right descriptive direction. It would also help to create complementary reference works; Lemmens and Wekker's proposal for a single 'dictionary/ grammar' (1991:13) need not be the only way.

## Notes

1. Note in passing that, due to this interplay; there may be conflicting information. It also represents a major difficulty for those users who use different dictionaries on occasion. It is their task to see if, and to what extent, different dictionaries agree on particular points.
2. Note the error with sense 2 of *start*, encoded as "It" ("It started to rain"), similar to "He hesitated to ask". "It" is to be understood as an intransitive verb with a to-infinitive clause functioning as an adjunct (cf. p 1559).
3. Treatment is extremely heterogeneous within and between reference materials, even if one leaves aside accidental differences, i.e. as to where usage and other differences are dealt with and what counts as a usage problem. Firstly, all dictionaries have much longer entries for *start*, i.e. about one column, than *begin*. *Begin* has about half a column in OALD, 15 lines in LDOCE, and three quarters of a column in CCELD. Secondly, the treatment of both verbs within the same dictionary varies greatly, despite their semantic similarities. For instance, OALD starts with the notional sense of *start*, while others begin with the aspectual meaning, etc.

## References

*Collins COBUILD English Grammar*, 1990. London: HarperCollins Publishers.

*Collins COBUILD English Language Dictionary*, 1987. London: HarperCollins Publishers.

Dixon, R.M.W. 1991. *A New Approach to English Grammar, on Semantic Principles*. Oxford: Clarendon Press.

Gramley, S., 1988. Infinitive and *ing* constructions as verb complements, in: W.-D. Bald, ed., *Kernprobleme der englischen Grammatik*, Munchen: Langenscheidt-Longman. 67-90.

Hausmann, F.J., 1989. COBUILD and LDOCE II. A comparative review. *International Journal of Lexicography* 2(1). 44-56.

Kjellmer, G., 1992. Grammatical or native-like? in: G Leitner, ed, *New Directions in English Language Corpora*. Berlin: Mouton de Gruyter. 329-344.

Legler, B., 1975. *Infinitiv und Gerundium nach den Verben BEGIN, CONTINUE, CEASE* [etc.] *in der englischen Allgemein- und Fachsprache der Gegenwart*. Dissertation an der Philosophische Fakultat der Universität Halle-Wittenberg.

Lemmens, M. and H. Wekker, 1991. On the relationship between lexis and grammar in English learners' dictionaries. *International Journal of Lexicography* 4(1). 1-14.

*Longman Dictionary of Contemporary English*, 1987. London: Longman.

*Oxford Advanced Learner's Dictionary*, 1990. Oxford: Oxford University Press.

Palmer, F.R., 1987. *The English Verb*. London: Longman.

Perlmutter, D., 1979. The two verbs *begin*, in: R. Jacobs and P. Rosenbaum, eds., *Readings in English Transformational Grammar*. Waltham, Mass., Ginn and Company. 107-119.

Quirk, R., *et al.*, 1985. *A Comprehensive Grammar of the English Language*. London: Longman.

Sinclair, J., 1986. First throw away your evidence, in: G. Leitner, ed., *The English Reference Grammar*. Tübingen: Niemeyer.

Sinclair, J. ed., 1987a *Looking up: An Account of the COBUILD Project in Lexical Computing*. London: HarperCollins Publishers.

Sinclair, J., 1987b. Grammar in the dictionary, in: J. Sinclair, ed., *Looking up: An Account of the COBUILD Project in Lexical Computing*. London: HarperCollins Publishers. 104-115.

# HUMAN AND INHUMAN GEOGRAPHY: ON THE COMPUTER-ASSISTED ANALYSIS OF LONG TEXTS

*Michael Stubbs and Andrea Gerbig*
FB2 Anglistik, Universität Trier

*"Human beings make their own history."*
(Karl Marx 1852.)

*"Human beings make their own geography."*
(Anthony Giddens 1984:363.)

## 1. LANGUAGE, TEXT AND MEANING

In an article published over 25 years ago, Sinclair (1965) points to major issues of scale, comparison and interpretation in text analysis:

> "Any stretch of language has meaning only as a sample of an enormously large body of text; it represents the results of a complicated selection process, and each selection has meaning by virtue of all the other selections which might have been made, but have been rejected."

The computer-assisted corpus work of Sinclair and others has shown how significant progress can be made on these difficult questions.

We discuss here the lexis and syntax of one long text to illustrate methods of computer-assisted analysis and the kinds of interpretation which this can support. A "long text" might be a complete novel or the transcript of an hour's conversation. Our example is a school geography textbook of about 80,000 words. So major linguistic descriptive questions are:

> What patterns of meaning exist across long texts, and what methods are available for describing and interpreting them?

We will discuss mainly how change, cause and effect, and agency (inevitable topics for a geography book) are represented. Such features can be fully interpreted only relative to their occurrence in other texts. So a further major question is:

> How can an individual text be located in diatypic space relative to other texts, text types and text corpora?

This comparative question is the superordinate one. Although we will make some descriptive progress, it will be apparent that not enough is known about which linguistic features might be significant, and that relevant comparative data are available only in fragmentary form.

We need an individual text to focus the discussion. From the point of view of the methodology of computer-assisted analysis, the choice of a particular geography book is arbitrary. But as an example of a text about the environment, the choice is not arbitrary: school textbooks and other discourses about the environment are important and massive uses of language.

### *1.1. Real texts, long texts and text corpora*

Despite substantial and increasing analysis both of linguistic corpora and also of real texts, methods of analysing individual *long* texts have received little discussion. In fact, despite much work in text analysis, as Phillips (1989: 8) ironically puts it:

> "Linguistics has traditionally been restricted to the investigation of the extent of language that can comfortably be accommodated on the average blackboard."

Most of the widely available major corpora themselves consist not of whole texts, but of text fragments, such as the 2,000 or 5,000 word samples in the LOB and LUND corpora. An exception is the COBUILD corpus, where the principle was followed of including complete texts. In addition, much corpus work is concerned with characteristics of the language as a whole (eg for lexicography) and not of individual texts.

Work in the very different tradition of critical linguistics (eg Fairclough 1989, Fowler 1991) analyses real spoken and written language in institutional contexts, but is largely based on short text fragments. Where analysis is based on longer texts or text corpora (eg Fairclough 1991 on speeches by a government minister, or Fowler 1991 on newspaper articles), the corpora are often not precisely specified, illustrative analyses are based on short extracts, and the corpora are not systematically searched for the phenomena under discussion. Indeed, since the corpora are not machine readable, such systematic searching could not in practice be done. Questions of whether analyses are exhaustive or replicable are therefore simply not raised[1].

So, a third question is:

> How can the intuitive insights of "critical linguistics" be developed and strengthened by computer assisted text analysis?

## 2. INITIAL OBSERVATIONS ON THE TEXTBOOK

Our text is a secondary school book on the physical and human geography of Britain, with sections on relief and climate, industries and transport, and regions, classified as urban, rural, industrial, recreational and "peripheral". (Scotland, for example, is "peripheral". A basic principle of text analysis is that all texts are written from a point of view!)

The layout is clear and attractive. On every double page spread there are maps, photographs, diagrams or charts, and questions. The book has relatively short sentences (circa 14 words) and often short sub-sections (circa 200 words). Each two-page spread is relatively autonomous, and teachers can use the book sequentially, by geographical region, or can use cross-references to take together, for example,

all the sections on rural or industrial areas. Such a layout already makes assumptions about pupil-readers with limited attention spans.

There are no references to any other books or articles in the text, no indications of where the authors obtained their information, whether books, reports, personal observations, or other geographers.

## 3. THE LINGUISTICS OF REPRESENTATION

Our questions are: How is the world talked about? How are relations between people and the physical world represented? For example, are physical and human processes talked about in the same ways? In some cases at least, this seems to be so. For example, (1) to (4) from the book use the same lexis and intransitive syntax to talk about processes which are purely natural and geological (1), non-natural but not under human control (2, 3), and the result of a human decision, probably by a small (but unidentified) group of people (4):

(1) a great variety of natural landscapes *have developed* from different underlying rocks
(2) a distinct urban landscape *developed*
(3) industrial cities *developed*
(4) the forest area *developed* as a tourist attraction.

Our topic is therefore the *linguistics of representation* (Fowler 1991), along with discussion of the interpretation of data on the frequency and distribution of textual features. How can such observations be more systematically based on textual evidence?

### *3.1. Conventions*

We use the following conventions:

- Examples are unmodified citations from the text. (A very few exceptions are explicitly signalled.) We follow the principles that "there is no justification for inventing examples" (Sinclair ed 1990: xi), and that examples are "not to be tampered with" (Sinclair 1991: 5).
- The complete sentence is not always cited; in such cases no marker of incompletion is given since it should be obvious from the context where the sentence is incomplete.
- Round brackets enclose raw frequencies of occurrence.
- Upper case indicates a lemma. (This is interpreted widely to include all syntactic variants : eg DEVELOP includes both verb forms, eg *developed*, and noun forms, eg *developer, development*).

## 4. LEXICAL DENSITY

Our main aim is a comparative method which would allow us to locate texts in diatypic space, relative to other texts and text types. Some initial evidence can be gained from global lexical analysis.

For six 2,000 word samples from the beginning, middle and end of the book, we calculated one simple measure of richness of vocabulary: the lexical density (ie the

percentage of lexical or open class words, as opposed to grammatical or closed class words). See Ure 1971[2].

In the six samples, this measure varies slightly between 55 and 57 per cent. This figure is meaningless, however, unless we can compare it with the range of densities in texts of different types. We therefore calculated figures for the 500 samples of different genres in the LOB corpus of a million words of written English. In LOB, the range of densities varies from 38 to 60. The 500 texts samples are distributed symmetrically around a mean (and median) of 48. Three quarters of the LOB texts have lexical densities below 51. 98 per cent have densities below 56.

Now we can place the geography book relative to these figures. Its lexical density is high, relative even to learned and scientific articles and press reports written for adults. It is higher than all the LOB non-fiction texts, and higher than many academic articles. In fact, it is among the 2 or 3 per cent of the most dense LOB texts.

This statistic provides a simple and global, but objective, measure of the fact that the book contains a high percentage of fairly technical words, packed together in informationally dense syntax. And it provides a useful rough comparison of this text to others.

## 5. CHANGE, CAUSATION AND AGENCY

A central assumption in the linguistics of representation is that the same events can always be talked about in different ways. We therefore examined the *grammatical encoding* of human activities in the book. The language of the text mediates reality by representing things in some ways rather than others. So which means of representation are chosen and what do these choices mean?

An important topic for a book on human geography is change: development or decline. And a textbook must explain things: so how are cause and effect represented? The text is about both physical and human geography: so how are physical and human agency represented?

Change, causation and agency are complex semantic areas. We will discuss:

- *passives*: a feature of certain text types, discussed, almost to the point of truism, in many studies; but often discussed purely intuitively, without comparative corpus data
- *ergatives*: a linguistic feature less discussed in stylistic studies
- realizations of *subject nominal groups*

and more briefly:

- *-ing*-forms of nouns and verbs.

### *5.1. Representation of processes in passives*

One much commented feature of texts, from scientific articles to newspaper stories, is passivization. This is usually discussed formally (with reference to word order changes between active and passive, and to agent deletion), functionally (with reference to theme, end focus, etc), and ideologically (as a way of avoiding mention of agency).

Although passives are widely discussed, basic descriptive, comparative and interpretative points are often ignored. To this extent, Bell's criticism of much work in critical linguistics and semiotics is justified:

> "... most studies (have) leapt past the groundwork to premature conclusions about the significance of sometimes poorly described linguistic patterns." (1991: 215-16)

A valuable source of comparative statistics is Svartvik (1966), which is further interpreted by Halliday (1991). Svartvik studied spoken and written corpora, from the survey of English Usage, plus other texts of his own.

For our text, we have studied

- the frequency of passives
- the frequency of *by*-passives
- the types of agent in *by*-passives, whiz-deletions and nominals.

*5.2. Passives with and without agents*

Svartvik (1966) calculates the number of passive clauses per 1,000 words of running text. For his 320,000 word corpus of eight text types, this averaged 11.3, from 3.0 in advertising to 23.0 in science. Active is overall more likely than passive, but this probability varies between around 9 to 1 and around 3 to 1 in different text types.

Our text contains circa 20 passives per 1,000 words of running text. In comparison with Svartvik's corpus, our textbook has a high density of passives, only slightly lower than specialist scientific writing.

Leech and Svartvik (1975: 258) claim that "about 4 out of 5 English passive clauses have no agent" (with no clear definition of the data on which their claim is based). In our text, of all passives, more than 7 in 8 are agentless. In so far as this comparison is reliable, there is relatively less expression of agency in our book than in the language in general.

Passives are agentless even when small local activities are being referred to:

- ... tweeds *are still woven* from local wool in the Outer Hebrides. ... Peat *is still used as fuel*, and *can be seen* stacked outside single-storey crofthouses.

We now have two independent measures (lexical density and frequency of passives) in comparison with two different corpora. Both measures place the school textbook amongst scientific writing for adults.

*5.3. Agents after passives, etc*

We studied all cases of agents in a *by*-phrase, not only after passives (202), but also after whiz-deletions (45) and nominals (5), as in, respectively:

- the parks *were set up by* local authorities
- a low plateau *covered* by chalky boulder clay
- *investment* by the government

Very few (5) agents are identifiable individuals, a few more (22) are groups of people, a few more again (38) are organizations, as in, respectively:

- Sinclair Research was set up *by Sir Clive Sinclair*
- the closures were opposed *by the dockers*
- £43 million was loaned *by the EEC*

By far the largest group of agents (187) is "other". Clearly, in some cases (22), the whole point is that no human agency is involved:

- this type of rock has been changed *by heat and pressure*

But in other cases (165), human agency is made abstract or is not expressed:

- cows are milked by *machine*
- apples and pears are picked *by local casual labour*
- oil is brought ashore *by two methods*
- areas affected *by steel closures*

### *5.4. Relative frequency of passives: stylistic profiles*

We can display findings in a way which shows more clearly the possibility of comparative studies, but also shows where comparative data are simply missing, due to the lack of such studies (or due to our ignorance of such studies). We will use these conventions:

L = average occurrence in the language as a whole (as represented, however imperfectly, by corpora)

G = average occurrence in the genre: here textbooks

T = actual occurrence in text under study

? = figures unknown!

| | L | G | T | |
|---|---|---|---|---|
| passive | 11.3 | ? | 20 | per 1,000 words. |

This says: in English as a whole, passives occur at an average frequency of 11.3 per 1,000 words of running text (according to Svartvik/Halliday). We do not know of relevant statistics for a corpus of school textbooks. In the present text, the frequency is 20 per 1,000 words.

Passive is an entry condition to other systemic choices, which can be shown with their probability:

| | L | G | T |
|---|---|---|---|
| agentless | 4 | ? | 7.5 |
| with agent | 1 | ? | 1 |

The ratio of 4 to 1 is from Leech and Svartvik (1975).

### *5.5. The grammar of individual words: semantic field of "change"*

The topic of "change" is reflected in the high text frequency of several verbs of change. In addition, the grammar of individual words in this semantic field gives evidence on how the causes and agents of change are represented, and also on the metaphors used to represent change.

### *5.6. Representation of processes in ergative verbs of change*

Sinclair (ed. 1990: 155ff) points out that many verbs of change are ergative. These are verbs which can be transitive or intransitive; and which allow the same nominal group as object in transitive clauses, and as subject in intransitive clauses[3].

For example (using brief but unmodified instances from the text), in

- several firms *have closed their factories* (transitive)
- factories *have closed* (intransitive)

*factories* is object and subject respectively: but in either case the closing happens to the factories. The structure is one way of avoiding mention of agency. Other structures can also be used with the same lemmas, eg:

- ... factories *have been closed* (passive)
- ... caused *the closure* of many factories (nominal)

Halliday (1985: 144ff) points out that after a passive, it is possible to ask "who by?" or "what by?". This is not possible after the intransitive option. He analyses ergativity semantically as a pattern of transitivity which is based on one variable of causation: whether the process is represented as being caused from without, or from within, as self-caused: He claims that the majority of verbs in common use in English are ergative[4], and that this pattern has come to prominence as part of a far-reaching process in modern English, which has left the transitivity system in a particularly unstable state.

Ergative verbs of change include: *change*, *close*, *decrease*, *develop*, *grow*, *improve*, *increase*. (Fillmore 1969 and Sinclair ed 1990 provide further examples.) We studied the text frequency of various grammatical options for these verbs.

For example, CLOSE occurs 80 times. All these occurrences collocate with such nominals as *factories*, *plants*, *works*, *firms*, *mines*, *mills*, *docks*, *schools*, and *railway lines*. Occurrences are: intransitive (41); nominals (24, all agentless); passives (9, all agentless); transitive (6). These various grammatical choices have different implications for the expression or omission of causation and agency. In the few transitives, the agent is never an individual, always an organization (*ICI*, *BSC*, etc) or a more abstract metaphor, eg:

- the "Beeching axe" *closed* hundreds of lines

The clear predominance of intransitives, nominals and passives leaves agency inexplicit.

There is a similar pattern with DEVELOP (276). The corresponding figures are: nominals (189), intransitives (51), transitives (19, two with human subjects) and passives (11, all agentless); eg, respectively:

- the stages of *development* for western economies
- air links *have developed* between the mainland and islands
- Birmingham *developed* a jewellery quarter
- Aviemore *has been developed* as an all-year centre

A summary for several ergative verbs is as follows:

| | intrans | passive | nominal | trans | absolute frequency |
|---|---|---|---|---|---|
| | percentages | | | | |
| CHANGE | 27 | 3 | 67 | 3 | 132 |
| CLOSE | 51 | 11 | 30 | 8 | 80 |
| DEVELOP | 19 | 4 | 70 | 7 | 270 |
| IMPROVE | 8 | 22 | 32 | 38 | 72 |
| INCREASE | 55 | 1 | 35 | 9 | 98 |

(In this table, adjectival uses are omitted, except that noun modifiers are counted as nominal.)

Although there are different distributions with different verbs, the total of the first three columns, which leave causation unexpressed, always outnumbers column 4.

*5.7. Comparison of ergative verbs across texts*

Again however, without comparative data from other texts, we have no way of interpreting such figures. We do not know whether such distributions are typical of English or whether they represent a particular tendency of this text to leave agency unexpressed. We have therefore compared this text with others.

For example, we have studied a school book on ozone depletion and protection, prepared by an Australian environmentalist group. Compared with the geography book, there are fewer passives (13.5 per 1,000 words), the same percentage of *by*-passives, but fewer abstract agents; and with ergative verbs, there are more transitive and fewer intransitive uses. This is consistent with the explicitly environmentalist stance of the Australian book: agency and responsibility are more explicitly expressed.

We have also studied the text frequencies of intransitive, passive and transitive structures used with ergative verbs across a corpus which represents (however imperfectly) written English, the LOB corpus of 1 million words of various genres: newspaper articles, bureaucratic reports, academic writing, fiction, etc.

The following are the summed figures for verbal structures of the five lemmas CHANGE, CLOSE, DEVELOP, IMPROVE, INCREASE, for the geography book (text 1), the corpus (LOB) and the environmentalist school book (text 2):

| | intransitive | passive | transitive | absolute frequency |
|---|---|---|---|---|
| | | percentages | | |
| Text 1 | 65 | 14 | 21 | 294 |
| LOB | 37 | 20 | 43 | 733 |
| Text 2 | 36 | 12 | 52 | 91 |

These five words (as is predictable from the topic) are relatively much more frequent (circa 5 times) in text 1 than in the corpus.

It is clear from the percentages that there are differences amongst the two geography books and the corpus. The relative percentages of transitive and intransitive are reversed in the two school books. And this reversal is in the intuitively expected direction: the "environmentalist" text expresses causation and agency more frequently through more frequent transitive constructions. In addition, the environmentalist text is much closer to the norm for the language, as represented by the corpus.

(These patterns are so striking that statistical tests of significance are probably not necessary. However, a check with the chi-squared test shows the following. The differences between the two books, and also between text 1 and the corpus, are statistically highly significant, with a probability of less than 0.001 that the differences have occurred by chance. The difference between text 2 and the corpus is not significant at the 0.1 level.)

This provides a very clear example of ideological point of view being conveyed by grammatical patterns. Such patterns could only be studied on the basis of textual

data, and without computer assistance such study across more than very short texts would be impractical or impossible.

In work in progress, we are doing a more extensive study of ergative verbs across other corpora, in order to check (amongst other things) Halliday's (1976, 1985) claims about the overall frequency and significance of this verb class.

### *5.8. Non-ergative verbs of change*

Not all verbs of change are ergative, and other verbs in this semantic field can also provide insight into how processes are represented. We studied lemmas such as: RISE, GAIN, DECAY, DECLINE, DROP, FALL, LOSE, NEGLECT, REDUCE.

For example, LOSE occurs 56 times. The most frequent collocations are with *job(s)* (28), with *trade*, *market*, *money*, etc (10), and with *population* (6), eg respectively:

- 6,000 men *lost* their jobs
- British Rail *lost* passenger and freight revenue
- the urban cores *are losing* population

It can be difficult to see that such ways of talking embody metaphors. But people do not "lose jobs" as they might lose money, by carelessness or gambling. Other people sack them. When the collocations are further nominalized (eg *job losses* (8)), the process is represented as a state.

The lemma DECLINE occurs 203 times. Our analysis confirms Sinclair's (1991: 46ff) analysis of the lemma from the COBUILD corpus. On his concordance evidence, there is no clear distinction between the senses "decrease (in quantity)" and "deteriorate (in quality)". In our data, collocations with *jobs*, *population* and numbers tend towards the "decrease" sense. Collocations with *suffer*, *industry*, *town* and *in* (*in decline*) tend towards the "deteriorate" sense. But many are indeterminate. The most common collocates (within a span of -10 and +10) are *industry/ies* or mention of specific industries (95), numbers (38), *population* (22), *employment* or *job* (19), *town*, *city*, etc (8), *suffer* (6). There are also a large number of collocates from the semantic field of "change" itself. Eg:

- the population of the Valleys continues to *decline*
- powers to stop decay and *decline* in inner city ... areas

Sinclair (1991: 50) notes that the collocate *Britain* tends towards the "deteriorate" sense!

Halliday (1990) discusses areas of English grammar which encode assumptions that growth is good, and small is not beautiful. The lack of a clear distinction between the two senses of DECLINE is one example of this. At this point in the lexis, the language does not distinguish between decline in quantity and in quality. Semantic distinctions can be ideologically significant.

### *5.9. Representation of human agency in sentence initial position*

Further data on the representation of human agency can be collected by studying the surface grammatical subjects of sentences. We studied the *subject nominal group* in declarative sentences, usually (in this book) the first constituent of the clause. Where the nominal group is anaphoric (eg *It*, *This*, *They*), we categorize it

according to its anaphoric referent. (Other sentence initial items, eg *There* (*is*), are ignored here.)

As a first generalization, only a small minority of subject nominal groups name the agent or logical subject of the verb. Very few (17) subject nominal groups name individual and/or identifiable people:

- *The port's managing director* puts the success ... down to
- *William Lever* built one of the world's first industrial villages

More (82) name individual or identifiable organizations with an official or legislative existence:

- *Hoechst Chemicals* have a research centre there
- *Local authorities* can plan for schools, housing

(Cf *the London Docklands Development Corporation, the Welsh National Party, Plaid Cymru, the EEC countries.*)

More again (200) refer to groups of people, but do not uniquely identify the referent:

- *Young people* moved away to seek work
- *The visitor* can fly to the regional airport

(Cf *Asians, Britain's ethnic minority population, tenants, Londoners, millions of tourists, opponents of the policy.*)

However, the largest number give no lexical and grammatical expression to what is logically an underlying human agent. This category includes:

cases (83) where the *head noun* is abstract, but implies groups of people (eg *firm, growth industries, workforce*):

- *Mines* would employ up to 500 people
- *The farming workforce* has declined so greatly that
- *The ethnic structure of the population* has changed

and cases (216) where an abstract noun refers to human activities or their results:

- *improvement* underway includes the Docklands Relief Road
- A *comparison* with any Third World country illustrates
- A *report* on the Brixton (London) riots in 1981 identified
- The *loss of jobs* at Devonport's led to Plymouth being
- *Out-migration* continues, however

(Cf *exports, nationalization, policy, production, project, schemes, survey, target, unemployment, unrest* and *violence.*)

Such cases include gerunds (*-ing*-nouns):

- Early *fertilising* may be necessary
- The average *spending* of overseas delegates in 1978 was ...

Other nominal groups (20) refer to concrete entities, but, again, grammatically disguise human activity:

- *The inner docks* have closed
- *Only 10,000 vehicles* a day crossed the bridge
- *Oil rigs* have drilled test wells

Many of the grammatical metaphors involved in such constructions seem entirely "natural". And it is probably just such examples which are most significant ideologically. It is with reference to such examples that we might say (in Marxist terms) that the actual labour of real working people is hidden. The material processes (in both a Marxist and a Hallidayan sense) of social life are given abstract representations by the grammatical choices, presumably largely unconscious. Throughout the book, both people (*labour*, *workforce*, *structure of the population*) and people's actions (*report*, *survey*) are encoded statively as abstract nouns.

In other examples (59), involving purely physical geography, no human agency could be involved:

- The peat fen was formed by the slow decay of vegetation
- The Skiddaw slates produce more rounded mountains

Note that in these cases agency (cf. *produces*) may be attributed to natural objects.

### *5.10. Other lexical and syntactic markers of change*

There are many other grammatical ways of representing things as unchanging (eg nominalizations, and stative constructions such as *there are ...*), or as changing (eg *-ing* forms of verbs).

There are circa 150 *-ing* forms of verbs, such as:

- the technical college has closed and shops *are closing*
- producers such as ICI *are increasing* their range

As would be expected, many of these (circa 50 per cent) co-occur with lexical indications of time or change, within a collocational span of about 20 words to left and right (ie -10 and +10):

- *today*, coniferous trees *are being planted*
- ... *by 1980*, and BSC *was losing* over £1 million a day

Although such constructions indicate change, very few (about 6 on a generous interpretation) collocate with lexical indications of cause or explanation for the change:

- Britain *was facing* an energy crisis *because* timber was running out
- the traditional farm *is becoming* rare. *As a result, ...*
- London *was growing* faster than the rest of Britain. *The causes* were...

### *5.11. Interpretative summary*

The cumulative effect of various syntactic choices

- the high frequency of passives
- the high frequency of intransitive uses of ergative verbs

(both relative to other text types)

- the low percentage of agents
- the very low percentage of human agents
- the very low frequency of collocations between *-ing*-verbs and lexical expressions of causation

is to de-emphasize causation, and particularly human agency, even in aspects of human and economic geography. Our analyses show the representation of a world where human beings are largely absent as responsible agents, where processes take place spontaneously or are caused by other abstract processes.

*5.12. Progressive focussing: '-ing'-nouns*

It is impossible to study everything, but once overall textual patterns begin to emerge, it is possible to make predictions, and a methodological strategy is *progressive focussing*. Both in the language as a whole and in individual texts, macro patterns are reproduced in micro details.

For example, we commented above on abstract nominals. One common (622) type of abstract noun is the gerund (*-ing*-noun), which is used to refer to activities and processes, and is functionally related to nominalization. One relevant lemma is HOUSE.

This occurs 100 times as a nominal. Overall, *housing* is more frequent (57) than *house* or *houses* (43). In some cases, both forms occur: *terraced houses* (5) and *terraced housing* (3). In other cases, there is no choice possible: *public house* is not the same as *public housing*! But overall, the tendency is towards the more abstract representation, as in the first rather than the second example below:

- employment, housing and education
- factories, ... houses and schools.

## 6. COMPREHENSIVE TEXT ANALYSIS?

The rest of the article discusses problems which we have so far only touched on.

Crystal (1991), with reference to language profiles in clinical linguistics, points out that profiling is in principle comprehensive. But he discusses problems in constructing such profiles: the very large number of variables involved, the imprecision with which some variables can be defined, and the lack of normative comparative data.

The possibility of the comprehensive analysis of texts is explicitly posed by Barthes (1968, in 1986: 143):

> "... if analysis seeks to be exhaustive (and what would any method be worth which did not account for the totality of its object, ie in this case, of the entire surface of the narrative fabric?) ... Is everything in narrative significant, and if not, ... what is the significance of this insignificance?"

Yet Barthes also immediately undermines the possibility of comprehensive text analysis. Post-Barthes, we all believe in intertextuality. No text is truly autonomous: a school geography book is related to other school books, to texts in academic geography, and to other discourse about the environment. And no text can be definitively deciphered: the tissue of signs can be unravelled, but there is no end to the fabric of intertextuality, and no definitive interpretation of a text.

As often, long before more currently fashionable formulations, Firth (1950: 44) provided a succinct warning:

> "The statement of meaning cannot be achieved by one analysis at one level, in one fell swoop."

The problem is discussed within linguistic stylistics: Fowler (1975) provides a good summary. For a text which is longer than a few lines, an analysis might conceivably be exhaustive at one level of analysis and/or delicacy. But this risks vacuity: what about everything else? Anyway, the concept of a "total analysis" is model dependent. It would include too much detail, in which everything would be reduced to banality. And it would be merely taxonomic, reordering and describing the text, without explaining it.

### *6.1. Interpretation*

In more recent work in "critical linguistics", Fowler (1991) provides detailed, plausible interpretations of linguistic features of British newspaper articles, but then appears to lose confidence in his procedures: "there is no constant relationship between linguistic structure and its semiotic significance", and everything depends on an analyst who "must be very well informed, and must have learned by experience" (p.90). Taken together, so seriously do these two comments undermine claims of reliable interpretation, that it is difficult to believe that Fowler is being entirely genuine in his hedges. Why would he have written the book at all, if he thinks so poorly of its methods?

Although Barthes' point about the impossibility of a definitive interpretation of a text is widely accepted, there is no clear linguistic theory of the fuzziness of textual interpretations.

It is clear that an interpretation does not spring mechanically from statistics. There is a category shift as we make inferences from ways of talking to ways of thinking. But quantitative data provide evidence for otherwise much more subjective interpretations of the text.

One plausible, but perhaps over-simple, assumption is that frequency of occurrence itself conveys meaning. One can isolate small units (syntactic structures, lexical categories, etc), whose frequent use has an additive effect on the meaning of a long text. This view has well known problems. The criteria for selecting the units may be unclear: they may be the easiest to identify, not the most relevant. Second, such units may not have the same meaning each time they occur: see below on passives. Third, selecting small units may ignore larger structural discourse organization.

The argument remains that if certain ways of talking about the world are systematically selected, to the exclusion of other ways, then this coding orientation (Halliday 1990, Bernstein 1990) will have a cumulative effect on ways of talking and thinking about the world. If, for example, grammatical metaphors such as

- The area lost population every decade between 1851 and 1961 (example from the text)

are systematically preferred to

- People left (or were forced to leave) the area every decade between 1851 and 1961 (invented alternative)

then it is at least plausible that these codings will have an effect on ways of thinking about problems of depopulation.

Further, if various different patterns point in the same direction, we can have more confidence in the interpretation. Thus, our analysis shows a tendency to leave causality unexpressed in several independent syntactic and lexical patterns.

*6.2. Interpretation, comparison and scale*

Questions of scale create both opportunities and problems. As Sinclair (1991: 100) says:

> "The language looks rather different when you look at a lot of it at once."

So, how are frequency and distribution to be interpreted? What does it *mean* if dozens or hundreds of examples of a lexical or syntactic pattern occur in a long text? It must be *relative* textual frequency which is of primary interest. Relative to other texts or text types, a feature might be used only in the variety under study, or with average frequency, or with above or below average frequency, or never. (Crystal 1991.)

It is uncontentious that cumulative frequency is a realization of register or text type. However, possibly the most serious limitation at present on a semantic interpretation of textual frequency data is the lack of normalized comparative statistics from a range of text types. The literature is full of claims about stylistic features which are "common" or "dominant" in certain text types. But such claims generally rely on intuition, which is notoriously unreliable on frequency. (Biber 1988 does provide just such detailed comparisons for two corpora.)

Finally, because of the amount of data (eg tens of thousands of words in a "long" text, and possibly tens or hundreds of millions in a corpus), a human observer simply cannot remember the data or directly observe the patterns. Only computers can provide the necessary indirect methods of observation via searching and pattern recognition.

So what does it mean to claim that there are patterns in the data which are not directly observable by human beings? What claims are being made about unconscious or subliminal recognition of patterns? Such patterns must be part of linguistic competence, since they consistently reproduce differences between genres. But we are left with particularly difficult interpretative questions, involving the relations between patterns of language (only indirectly observable) and patterns of thought (at best semi-conscious).

These same problems have long arisen in literary criticism. In computational studies, we are at least forced to make explicit the patterns which we want the machine to find.

*6.3 Interpreting passives*

We cited figures above for the proportion of actives to passives in various text types. However, it is widely agreed that linguistic features have no single interpretation. Passives, for example, have a thematic function of moving information to different places in the clause. This also allows the agent to be omitted. But this may be for various reasons: because the information is obvious, or unknown or irrelevant, or in order (consciously or not) to be vague about or hide the information.

The passive is fully productive syntactically and semantically. But some verbs never occur in the passive (eg *become*), whilst other always do (eg *born, reputed*). And, based on corpus data, Sinclair (ed 1990: 407) gives many examples of verbs which could be active, but generally aren't: eg *be baffled, be rumoured.* Thus active-passive choices are not available on all verbs. Furthermore, there may be

little real syntactic choice, in so far as passives are simply conventional in some text types. Historically, the passive was explicitly motivated in scientific writing, to signal the unimportance of the individual experimenter (Swales 1990). But nowadays its use may simply be conventional. These considerations complicate inter-text comparisons.

In addition, different studies are usually not directly comparable due to different definitions of linguistic features: eg does "passive" include whiz-deletions (*money invested by the government*)?

*6.4 Where do the features come from?*

We cannot analyse all the linguistic features of a long text. So where do the analysed features come from? This can only be from a mixture of intuition and published analyses, including published lists of such features (Crystal 1991). (There is no purely inductive data-driven description.)

It is known from many studies, of various text types, which linguistic structures are likely to be relevant to an analysis of factual writing. From the point of view of the linguistics of representation, Fairclough (1989, 1991), Fowler (1991) and Myers (1992) provide lists of linguistic features which are likely to be "particularly worth looking at" (Fowler 1991: 77). Such lists are plausible and based on detailed illustration and argument. However, they rely on an inexplicit accumulation of experience, are not formalized, and do not therefore distinguish between textual features which are easy, difficult or impossible to find with computer methods. Intuition is also likely to be of only limited help in identifying the clusters of features which co-occur in text types (Biber 1988).

We are certainly not starting from scratch in a study of school textbooks. Wignell *et al* (1987) discuss in detail the types and functions of technical lexis in school geography books for observing, ordering and explaining the experiential world: this could be the basis for a more detailed study. There are valuable "critical linguistic" analyses of sections of school geography books (Kress 1985) and other writing about the environment (Martin 1985). But this work analyses only short fragments of text, not cumulative patterns across whole books.

A criticism we anticipate is that we have only confirmed what is already known from other analyses. We have provided more illustrations, but textbooks are well known to use lots of passives, abstract words, etc. Have we fallen into the trap of merely confirming conventional wisdom by selective investigations? Linguists may think that English is well described and not see new things even when confronted with corpus evidence. Of crucial importance in avoiding this danger is to develop a more comprehensive and formalized list of features for computer assisted study. In general, we must beware of assuming that we already know which features are relevant.

*6.5 Methodological points: computer text searches*

Unlike the study of isolated and invented sentences in much recent linguistics, textual studies allow linguistic features to be studied in real contexts.

However, an important danger is counting only what is easy to count. With the pattern-matching facilities of concordance programs, it is relatively easy to identify individual words and phrases, lemmatized or not; closed categories of words (eg modals); some larger, less well defined, word categories (eg ergatives); and some

syntactic structures (eg passives). These are all relevant to studying the balance of information in clauses, representation of agency and causality, interpersonal and ideational meanings, the authorial stance towards the information presented, etc.

But some categories are easier to identify exhaustively than others. Our analysis is restricted mainly to surface forms identifiable with a concordance program. In addition, the input data were raw text, not grammatically tagged or parsed, and this places limitations on what can be found. For example, noun-noun constructions (eg *railway lines, factory closure, steel industry, job losses*) are frequent, but cannot be computationally identified from untagged text. Note however Sinclair's (1991: 21, 29, 117) "clean text" policy, and his warning about the analytic categories which have to be accepted along with a tagged corpus. He emphasises "the strength of patterning which emerges from the rawest of unprocessed data". Halliday (1991: 34) warns, conversely, of "the familiar catch that what is easy to recognise is usually too trivial to be worth recognising". But it is easy to underestimate what can be done with concordances of untagged text. For example, we identified sentence initial units by the trivial procedure of doing a concordance on full stops and question marks. This prints all sentence beginnings. These can be further ordered in various ways and it is then simple to inspect subject nominal groups as above) for characteristics, such as grammatical metaphor, which are not directly identifiable.

Concordances are often seen (Leech 1991: 19) as useful only for studying individual words or phrases. However, even with untagged text (*contra* Leech), the pattern matching of concordance programs can identify important syntactic patterns. For example, if the matching allows alternatives and wildcards, and allows lemmas to be built up, then passives with regular verbs can be identified with a high degree of accuracy, and with irregular verbs a small amount of simple post-editing can weed out wrong matches.

Search algorithms can always be refined. This is time-consuming, and a balance must be found between being certain of identifying every example in the data, and being contented with identifying more examples than could possibly be identified without computer assistance (with those examples missed being assumed to be infrequent, random and therefore not significant.) Other patterns may simply be unidentifiable in untagged text. But one does not abandon a very powerful observation method because it is not perfect.

The machine will never do all the analysis: but it can retrieve and display data in ways which allow the human to see different patterns[5].

### *6.6. Comparative analysis and software tools*

Any analysis is partial. For example, we have regarded the text as homogeneous, and have not studied any variation *within* the text. Further, we have studied relations between this text and others only briefly, and have pointed to the lack of relevant comparative statistics.

Over the next few years a wider range of software tools will certainly become available for corpus analysis. It is already possible to envisage an integrated software environment, in which a suite of programs could analyse different aspects of texts. Perhaps someone somewhere is developing such an environment, integrated and well documented. One would sit at the terminal, call up the text, and have access to various information and comparison routines:

- a concordance program with powerful pattern matching (as available in languages such as SNOBOL)
- programs to calculate word frequencies
- word lists (eg grammatical words, core vocabulary, sub-technical vocabulary) and programs to compare text against word list
- normalized statistics (eg type-token ratio, lexical density, proportion of active to passive clauses) for different genres
- grammatical tagging and parsing programs
- various text corpora, defined on various dimensions, for other comparisons.

Such software already exists, but only some is publicly available and robust. Some, such as our software for lexical density calculations, is homegrown. Even some of the major tagging and parsing programs, frequently referred to in the literature, are not currently robust enough for free distribution and use on different hardware.

### *6.7. Institutional analysis*

We have discussed the language of a textbook, but not its use in school. It is a truism, post-Barthes, that the meaning of a book is not merely "in the text": books are re-written with each re-reading. It is also a truism that there are no teacher-proof teaching materials. Teachers mediate textbooks, perhaps emphasizing their factual status or, alternatively, offering different interpretations. Textbooks can be read aloud or silently, memorized or discussed in small groups. This all requires ethnographic methods to study the external institutional context and how texts are received and consumed. (Stubbs 1992.)

It is known that the recall of broadcast news is very poor: recall is often as little as 5 per cent and rarely exceeds 30 per cent even on immediate questioning (Bell 1991: 232). In addition, for school pupils, the effects of a single geography textbook should not be exaggerated: it is just one more item in a constantly changing and fragmented school day, which has to compete with messages from a range of semiotic systems.

However, for the vast majority of people, almost all their knowledge of environmental issues comes from the media (Bell 1991: 239). The ozone layer and the greenhouse effect cannot normally be directly observed: these aspects of reality can only be linguistically constructed. And for most pupils, their knowledge of areas of the world beyond their local community can come only from representations in books, on television, and so on. In addition, textbooks are designed to be studied, learned and reproduced.

We conclude with more general implications of the kind of study proposed.

## 7. SCHOOL TEXTBOOKS AS GENRE

One of our underlying themes is the need to evaluate text analysis in terms of comprehensiveness, systematicness, reliability, comparability and replicability.

Now, evaluative criteria for an analysis make sense only with respect to some analytic purpose. For a school textbook, this might be to estimate its "readability" for pupils of a certain age, or its freedom from sexist or ethnocentric bias. All

analysis involves selection, and this selection must take into account the purpose for which the text is written and the purpose for which the analysis is being done. As Labov (1970) says, once the linguist knows what to count, the problem is practically solved.

Texts are written for many purposes: fame and fortune are just two. However, an analysis of a school textbook must take into account features of the genre. Textbooks are institutionally sanctioned versions of knowledge: what is believed to be worth passing on to pupils. They are one product of the cultural selections embodied in all curricula. They are factual writing, but "there are no brute facts" (Firth 1957: 29). Factual writing interprets the world (Martin 1985). English has ways of presenting information as factual in its systems of modality and evidentiality (Stubbs 1986). We might say that geographers are interested in facts (eg how certain they are); that linguists are interested in factivity (how language expresses certainty, tentativeness, etc); and that teachers and pupils should be aware of the relation between facts and factivity. So, in any text, some things are selected for discussion; these things are encoded in language forms; this encoding mediates and constructs reality.

### *7.1. School textbooks in educational institutions*

We have taken a narrow linguistic route to the tip of a large educational iceberg. Schooling is a major way in which a culture communicates with pupils, and textbooks are a major medium for this communication. Education cannot be value-neutral: it is intended to form beliefs and transmit values. In this sense, pedagogic theory is a branch of the theory of communication. Education involves making a selection from the culture (Williams 1989), and we have discussed some aspects of the question: Why present this knowledge in this way?

Textbooks contain the codified knowledge of a field, and their linguistic expression has important implications for the sociology of knowledge. They are repositories of what we hold to be true. They are a massive genre, a primary medium of formal education, and despite movements in child-centred education, still a major source of formal expository academic prose for school pupils. Yet they are unrepresented in the influential Brown and LOB corpora of written American and British English.

Textbooks are founded on a paradox. On the one hand, written language allows (as Popper argues) statements to be explicitly formulated and more easily studied, and therefore encourages a critical view of knowledge. On the other hand, their characteristically authoritative stance (a discussion of ideas, but not of their source; and of facts, but not of the authors' attitudes to those facts) encourages an uncritical view of knowledge.

Barthes (1968: 27) argues that a serious contemporary problem is the problem of the transmission of knowledge. The great structures of economic alienation, he argues, have been largely revealed. But the structures of the alienation of knowledge have not. In what ways is knowledge separated from its producers (eg geographers) and appropriated by others (eg textbook writers)? What of the origins of the knowledge, the means by which it was produced, and the points of view from which it is represented? Are these acknowledged or hidden from its consumers, teachers and pupils?

## 8. HUMAN (AND INHUMAN) GEOGRAPHY

History has been a fashionable preoccupation of philosophers and social theorists, but geography has been relatively neglected as a domain of knowledge. Distance in time has always seemed to involve deep philosophical problems, whereas distance in space seems simple. Yet the social division of space is just as complex and central to understanding the world as the social division of time. (Foucault 1980, Giddens 1984.)[6]

Geography is one important case of how we talk about the world. It involves measurement, inventory and classification: geographers gather some of the information on materials, transport, populations, and so on, which governments and industrialists use for planning. Along with history, geography helps to create the discourse of nationalism: it concerns territories and boundaries, and regional and national identity. Geography straddles the natural and social sciences: it has physical and human aspects, and must deal with both ideational and interpersonal meanings.

Other discourse about the environment includes travel guides and travel writing of many kinds (particularly common on television). And quite apart from discourse explicitly about travel and foreign countries, spatial metaphors are common in discourse of all kinds (eg *make space for, the moral high ground, on familiar ground*).

Geography textbooks are of particular importance, since they explicitly attempt to teach children to talk about the world in certain ways. Their analysis therefore has implications for teaching the kind of critical reading which underlies, for example, the proposals (so inimical to the British Conservative government) of the Cox Report for language awareness and cultural analysis. (DES 1989, Stubbs 1989.)

## 9. INSTITUTIONAL LINGUISTICS

This article is part of a project on the relation of texts to text types and text corpora and to social institutions. (Stubbs 1992.) One focus of this larger project is the analysis of texts which are public and/or authoritative: textbooks, newspaper editorials, speeches of famous persons, encyclopedias, legal judgements and government reports. A second focus is texts about the environment: specifically, a corpus of scientific, legislative, press and public relations texts about the destruction of the ozone layer.

This project concerns the construction of a humanist linguistics. Texts, spoken and written, comprise much of the empirical foundation of society: they help to construct social reality. Textual analysis, as is very clear in "critical linguistics", is a perspective from which to observe the deep structure of society: it makes ideological structures tangible. But much of the deep-layered patterning is only indirectly observable in long texts and by using computer assisted methods. Sinclair's recent computational work (eg 1991) provides a vision of the unsuspected patterns of language, spectacular regularities with endless variation, not limited to what any individual can perceive or remember.

**Acknowledgements**

We are grateful to the following. Susanne Jarczok and Brigitte Grote for preparation of machine-readable texts and for preparing programs for calculating lexical density and related

statistics. Richard Alexander, Marlene Faber and participants at seminars at Edith Cowan University, Perth, Western Australia for critical comments on an earlier draft; and Simon & Schuster Education, Hemel Hempstead, UK, for permission to store the geography book for research purposes in computer readable form and to reproduce extracts. The book, *The British Isles*, is copyright 1984 Neil Punnett and Peter Webber, and published by Basil Blackwell Ltd.

## Notes

1. Phillips (1989) is a computer-assisted analysis of complete science textbooks (from the COBUILD corpus), but deals only with lexical patterns concerning readers' perceptions of what a book is "about".
2. The software (written by Brigitte Grote) reads in any text and any word list (here, a list of grammatical words), calculates the percentage of words from the list in the text, and prints out listed and non-listed words alphabetically and by frequency for more detailed study.
3. This is a complex and terminologically confused area. Fillmore (eg 1969) bases his arguments for deep grammatical case on such verbs: he uses the term "objective" for the function which is the subject of an intransitive verb and the object of the corresponding transitive verb. Halliday (1976) uses the term "affected" for Fillmore's "objective". Lakoff (1970: 33ff, 44ff) discusses relevant verbs as "inchoatives" and "causatives". Palmer (1974: 92ff) calls them "pseudo-passives". Quirk *et al* (1985) use this term for something different, and talk simply of transitive and intransitive verbs with reference to causality. Halliday (1976) talks of "middle" and "non-middle" clauses. (Again, Quirk *et al* 1985 use the term "middle" verb for something different.) Halliday (1976) also talks of "neutral" verbs, since it is not possible to decide which form is basic (*she poured the water out* or *the water poured out*). Fontenelle & Vanandroye (1989) provide a detailed definition of ergatives according to their semantic, syntactic and morphological properties.

   Lyons (1968: 350ff), Halliday (1985) and Sinclair (ed 1990) use the term "ergative". This has the merit of labelling an interesting class of verbs, although the term is usually used in work on linguistic typology to refer to morphologically marked ergativity in languages such as Basque or Dyirbal. English is clearly not ergative in this sense.
4. Our initial study of the 500 most frequent verbs in English throws doubt on this assertion: we will report on this more general study of ergatives elsewhere.
5. Other features, such as lexical sets which are formal markers of semantic fields, may be open to computational study, but only with complex statistical methods (Phillips 1989).
6. A version of this paper was presented in 1992 in the former East Berlin to a group of teachers from East and West. The social and political division of space is more obvious in some places than others!

## References

Aijmer, K & Altenberg, B eds (1991) *English Corpus Linguistics*. London: Longman.

Barthes, R (1968) The death of the author. In *The Rustle of Language*. Oxford: Blackwell. 1986.

Bell, A (1991) *The Language of News Media*. Oxford: Blackwell.

Bernstein, B (1990) *Class, Codes and Control*, Vol 4. London: Routledge.

Biber, D (1988) *Variation across Speech and Writing*. Cambridge: Cambridge University Press.

Crystal, D (1991) Stylistic profiling. In Aijmer & Altenberg eds 1991: 221-38.

DES (1989) *English for Ages 5 to 16.* (The Cox Report.) London: DES & Welsh Office.

Fairclough, N (1989) *Language and Power.* London: Longman.

Fairclough, N (1991) *Discourse Processes and Social Change.* Oxford: Polity.

Fillmore, C J (1969) Toward a modern theory of case. In D A Reibel & S A Shane eds *Modern Studies in English.* NJ: Prentice-Hall. 361-75.

Firth, J R (1950) Personality and language in society. *The Sociological Review*, XLII: 37-52.

Firth, J R (1957) A synopsis of linguistic theory, 1930-1955. In *Studies in Linguistic Analysis.* Special Volume of the Philological Society. 1-32.

Fontenelle, T & Vanandroye, J (1989) Retrieving ergative verbs from a lexical data base. *Dictionaries: Journal of the Dictionary Society of America*, 11: 11-39.

Foucault, M (1980) Questions on geography. In Gordon, C. ed *Power/Knowledge*, Brighton: Harvester.

Fowler, R (1975) Language and the reader: Shakespeare's sonnet 73. In *Style and Structure in Literature.* Oxford: Blackwell. 77-122.

Fowler, R (1991) *Language in the News.* London: Routledge.

Giddens, A (1984) *The Constitution of Society.* Cambridge: Polity.

Halliday, M A K (1976) Types of process. In G Kress, ed *Halliday: System and Function in Language.* London: Oxford University Press. 159-73.

Halliday, M A K (1985) *An Introduction to Functional Grammar.* London: Edward Arnold.

Halliday, M A K (1990) New ways of analysing meaning. Paper read to AILA Congress, Thessaloniki, Greece.

Halliday, M A K (1991) Corpus studies and probabilistic grammar. In Aijmer & Altenberg eds 1991:30-43.

Kress, G (1985) *Linguistic Processes in Sociocultural Practice.* Deakin University Press.

Labov, W (1970) The study of language in its social context. *Studium Generale*, 32, 1: 30-87.

Lakoff, G (1970) *Irregularity in Syntax.* NY: Holt, Rinehart & Winston.

Leech, G (1991) The state of the art in corpus linguistics. In Aijmer & Altenberg eds 1991: 8-29.

Leech, G & Svartvik, J (1975) *A Communicative Grammar of English.* London: Longman.

Lyons, J (1968) *Introduction to Theoretical Linguistics.* Cambridge: Cambridge University Press.

Martin, J (1985) *Factual Writing: Exploring and Challenging Social Reality.* Deakin University Press. Reprinted by Oxford University Press.

Marx, K (1852) Der 18te Brumaire des Louis Napoleon. In K Marx & F Engels *Werke.* Vol 8. Berlin: Dietz. 1960.

Myers, G (1992) Textbooks and the sociology of scientific knowledge. *English for Specific Purposes*, 11: 3-17.

Palmer, F (1974) *The English Verb.* London: Longman.

Phillips, M (1989) *Lexical Structure of Text.* Discourse Analysis Monograph 12. Birmingham: English Language Research, University of Birmingham.

Quirk, R *et al* (1985) *A Comprehensive Grammar of English.* London: Longman.

Sinclair, J M (1965) When is a poem like a sunset? *A Review of English Literature*, 6, 2: 76-91.

Sinclair, J M, ed (1987) *Collins COBUILD English Language Dictionary.* London: HarperCollins Publishers.

Sinclair, J M, ed (1990) *Collins COBUILD English Grammar.* London: HarperCollins Publishers.

Sinclair, J M, (1991) *Corpus, Concordance, Collocation.* Oxford: Oxford University Press.

Stubbs, M (1986) A matter of prolonged fieldwork: notes towards a modal grammar of English. *Applied Linguistics*, 7, 1: 1-25.

Stubbs, M (1989) The state of English in the English state: reflections on the Cox Report. *Language and Education*, 3, 4: 235-50.

Stubbs, M (1992) Institutional linguistics: language and institutions, linguistics and sociology. In M Pütz ed *Thirty Years of Linguistic Evolution*. Amsterdam: Benjamins. 189–211.

Svartvik, J (1966) *On Voice in the English Verb*. The Hague: Mouton.

Swales, J (1990) *Genre Analysis*. Cambridge: Cambridge University Press.

Ure, J (1971) Lexical density and register differentiation. In G E Perren & J L M Trim eds *Applications of Linguistics*. Cambridge: Cambridge University Press. 443-52.

Wignell, P, Martin, J R & Eggins, S (1987) The discourse of geography. *Working Papers in Linguistics*. Dept of Linguistics, University of Sydney.

Williams, R (1989) Hegemony and the selective tradition. In S De Castell, A Luke & C Luke, eds *Language, Authority and Criticism: Readings on the School Textbook*. London: Falmer. 56-60.

# ON BEGINNING THE STUDY OF FORENSIC TEXTS: CORPUS CONCORDANCE COLLOCATION

*Malcolm Coulthard*
University of Birmingham

## Introduction

Twenty five years ago Jan Svartvik published *The Evans Statements: A Case For Forensic Linguistics* in which he demonstrated that disputed parts of a series of statements, dictated to police officers by Timothy Evans and incriminating him in the murder of his wife, had a measurably different grammatical style from the uncontested parts and a new discipline was born. Initially its growth was slow; in unexpected places there appeared isolated articles in which the author, often a distinguished linguist, commented on disputed confessions or purported records of interaction, or evaluated the ability of ordinary people to understand legal language or the likelihood of a given suspect being the speaker recorded on a tape, or disputed the authenticity of non-native-like language attributed to immigrants or aboriginals.

However, there was no attempt to establish a discipline or even a methodology for forensic linguistics. In these early days the work was usually undertaken as an intellectual challenge and almost always required the creation, rather than simply the application, of a method of analysis. By contrast, in the past five years, there has been rapid growth in the frequency with which courts in a series of countries have called on linguists as expert witnesses, and, in consequence, there is now a developing methodology and a growing number of linguists who act as expert witnesses, a few even on a full time basis.

## The Task of the Forensic Linguist

Despite the exaggerated hopes of solicitors and clients, linguistic analysis can, obviously, say nothing about the truth of what is contained in a document; however, to their ultimate satisfaction, it is often possible to provide evidence in support of one of two conflicting claims about a document. In all my civil and criminal cases the dispute has been over whether a given person was the author of (part of) a given document – in the majority of cases the linguist is asked by the defence to support a claim that an accused is not the (sole) author of a document; much more rarely s/he may be asked by the prosecution to provide support for the more difficult claim that

a defendant *is* the author. In both types of case the linguist needs to analyse, or at the very least refer to, not simply the disputed document but also an appropriate small and case-specific corpus gathered for comparative purposes.

There are two practical pressures on the forensic linguist: firstly, at least in Britain, most of the work is paid for by legal aid and it is therefore not easy to get permission to spend more than 10 hours on a given case, whereas the same linguist, wearing her/his academic hat, would not hesitate to spend well over 100 hours writing a short article; secondly, the legal system will often constrain the linguist to work within hypotheses about authorship created by others, the validity and strength of which s/he may not be able or even allowed to evaluate – s/he may, for instance, be given an anonymous letter with copies of work by the suspect and be asked to say whether there is linguistic evidence to suggest that the suspect *could* have written the letter. Such a brief simplifies the linguist's task, but may lead her/him to help, unintentionally, in a miscarriage of justice, by ignoring the likelihood of other candidate authors.

On the two occasions I have been involved in such cases, I have first established that there are linguistic grounds to suggest that the suspect was a *candidate* author and then insisted, somewhat to the surprise of the commissioning solicitors, on a secondary corpus, of documents written by his colleagues, in order to establish that the linguistic similarities were not equally typical of the other candidates. However, this insistence slowed down the production of the final report, in one case by well over two months, and sometimes the linguist is called in literally days before a trial begins.

Obviously, the evidence the linguist gives in court will, in almost every case, be probabilistic – despite the claims of Morton (1991) it seems highly unlikely that any method of 'linguistic fingerprinting' will be developed in the foreseeable future. For this reason linguistic evidence is currently more usable by the defence, where the need is to show 'reasonable doubt', than by the prosecution, where the need is to show 'proof'.

The questions the forensic linguist must confront are: firstly, are there sufficient linguistic similarities between the authentic samples and the disputed text to make the suspect a candidate author and secondly, if so, how likely is it that the linguistic similarities are idiosyncratic; in other words is the suspect the actual author or just one of a series of potential authors? It is on the second count that the forensic linguist is currently in some difficulty because s/he has no way of quantifying the reliability – unlike the forensic phonetician who has, for example, population data for fundamental frequency ranges and can thus go into court and say with confidence, for instance, that the average pitch of the speaking voice of a given accused is not only very similar to that of the voice on the tape, but also that this pitch is unusual and only occurs in 1% of the voices of the male population. The forensic linguist does not yet have norms nor even, in most cases, corpora from which the necessary norms could be derived and is thus restricted in the degree of certainty s/he can attribute to her/his conclusions. The forensic linguist today is in a situation analogous to that of the lexicographer in 1980.

What I propose to do in this article is to exemplify some of the methods currently being used by forensic linguists and at the same time to discuss the ways in which a corpus linguistics approach could dramatically improve the quality and reliability of the methodology.

In what follows I will use only examples taken from real cases. Although I have changed the names of people, places and organisations, in order to maintain confidentiality, the texts are in all other respects genuine and unedited; in particular all spelling and punctuation mistakes are as in the originals and thus will not be indicated with (sic).

## 1. Using existing corpora

In an early case I was asked to examine the confession statement, (reproduced as Appendix 1), of Derek Bentley, the 19 year old who was found guilty of the murder of a policeman and hanged forty years ago, despite the fact that the shot was fired by his companion, the under-age Chris Craig and that he himself was in police custody at the time at which the policeman was shot. Bentley, who was illiterate and had a measured IQ which put him in the bottom 1% of the population, claimed at his trial that the police had 'helped' him with the statement, whereas all the police officers on oath denied this and asserted that the statement was an accurate 'verbatim' record of what Bentley had actually said. In producing my report, commissioned to form part of an appeal for a posthumous pardon sent to and later rejected by the Home Secretary, I focussed on a series of linguistic features only one of which I will discuss here.

### *"then"*

One of the marked features of Bentley's confession is the frequent use of the word "then" in its temporal meaning – 10 occurrences in 582 words. This may not, at first, seem at all remarkable given that Bentley is reporting a series of sequential events and that one of the obvious requirements of a witness statement is accuracy about place and time. However, a cursory glance at a series of other witness statements suggested that Bentley's usage was at the very least untypical, and thus a potential intrusion of a feature of policeman register, which is related to a professional concern with the accurate recording of temporal sequence. I therefore created two small corpora, one composed of three ordinary witness statements, one from a woman involved in the Bentley case itself and two from men involved in another unrelated case, totalling some 930 words of text, the other composed of statements by three police officers, two of whom were involved in the Bentley case and one in another unrelated case, totalling some 2270 words, (cf. Fox,1992).

The results were startling: whereas in the witness statements there is only one occurrence of "then" in 930 words, by contrast "then" occurs 29 times in the police officers' statements, that is on average once every 78 words. Thus Bentley's usage of temporal "then" groups his statement firmly with those produced by the police officers.

In this case I was fortunate in being able to check the representativeness of my 'ordinary witness' data against a reference corpus, the Corpus of Spoken English, a subset of the Birmingham Collection of English Texts, which, at the time, consisted of some 1.5 million running words collected from many different types of naturally occurring speech. "Then" in all its meanings proved to occur a mere 3,164 times in the whole corpus, that is, on average, only once every 500 words, which supported the representativeness of the witness data and the claimed specialness of the police and Bentley data.

What was perhaps even more striking about the Bentley statement, in fact, was the frequent postpositioning of the "then"s as can be seen in the two sample sentences below, selected from a total of 7 occurrences in the 580 word text:

Chris **then** jumped over and I followed.

Chris **then** climbed up the drainpipe to the roof and I followed.

This has an odd feel. Not only do ordinary speakers use "then" much less frequently than policemen, they also use it in a structurally different way: for instance in the BCET spoken data they use "then I" ten times more frequently than "I then"; indeed the structure "I then" occurs a mere 9 times in the whole of the spoken sample, in other words only once every 165,000 words. By contrast the phrase occurs three times in Bentley's short statement, once every 190 words, a frequency almost a thousand times greater than that instanced in the corpus. In addition this "I then" structure, as one might predict from the corpus data, does not occur at all in any of the three witness statements, whereas there are 9 occurrences in one single 980 word police statement, as many as in the entire 1.5 million word spoken corpus. Taken together the average occurrence in the three police statements is once every 119 words. In other words, the structure "I then" does appear to be a feature of policeman's register.

More generally, it is in fact the structure *Subject* (*Verb/Object*) followed by "*then*" which is typical of policeman's register; it occurs 26 times in the statements of the three officers and 7 times in Bentley's statement. Interestingly, Svartvik (1968:29-32) had made the same discovery but had not actually made it explicit because the analytical category he had used was 'clauses with mobile relator', with the gloss that "such clauses include *then* and *also*". What he did not emphasise was that in *all* the 23 examples of the category in his corpus the 'mobile relator' was in fact realised by "then".

When we turn to look at the record of the oral evidence given in court during the trial and choose one of the police officers at random we find him using the structure twice in successive sentences, "shot him *then* between the eyes" and "he was *then* charged". In Bentley's oral evidence there are also two occurrences of "then", but this time the "then"s occur in the normal preposed position: "and *then* the other people moved off", "and *then* we came back up". Even Mr Cassels, one of the defence barristers, who one might expect to be have been influenced by the forensic style, says "*Then* you".

Recently more work on a whole series of other police and witness statements has confirmed the use of postposed "then" to be the most reliable distinguisher of police register. However, the forensic linguist still needs a reference corpus of authentic police and witness statements in order to be able to make statistically valid statements in court and to protect himself against the suggestion by hostile cross-examiners that, as any reasonable person will agree, a corpus of general conversation is irrelevant for comparative and normative purposes, because the linguistic behaviour of witnesses and subjects must change when they are making statements under oath.

## 2. On Investigating Idiosyncrasy

There follow extracts from a typed anonymous letter; the italicisation is mine:

> I hope you appreciate that *i* am *enable* to give my true *idenitity* as this *wolud* ultimately jeopardize my position...

> Another issue *I* would like to *high light* is a young man John Smith, *Subsequantly* came to us. He had *quiet alot* of *vaulables*...
>
> ...residents *how* are *refered* to us have very *little material thing* in life. I feel his *right* should have been protected...
>
> ...both Peter Brown and Mary Green are well aware of these *situation* and have so far *deened* it unnecessary to *investegate* these *issus*.

In this case I was initially sent the anonymous letter and a sample of work, both typed and handwritten, produced by the suspect in the normal course of his work. I subsequently requested and received samples from a further nine employees for comparative purposes. As is evident from the above short extracts from the 1,600 word text there are some fascinating features, of which the four most significant seemed to be:

a) the typing oddity of using "I", "i" and "1" interchangeably for "I",

b) the occasional separation of two-morpheme words into two separate words: "high light", "with out" "in to", "some one", "stand by", "my self";

c) the major problem with the orthographic representation of vowels which in the spoken version are reduced to schwa: e>u, caturing= catering; u>e, enable = unable; a>e, except = accept; i>e, investegate = investigate;

d) past tense forms: has threaten(ed); diagnoice(d); he grab(bed) her; has not change(d).

Only one of the ten sampled authors displayed all these idiosyncrasies, and this allowed me to conclude that, on the basis of the sample I had analysed, there was only one candidate author. However, as we have at the moment no corpus of adult writing there is no principled way of saying which of the above kinds of mistake might be expected to occur more or less frequently in texts and thus which might be a more significant indicator of authorship. For instance, one of the other nine authors sampled in this case also had a tendency to split two-morpheme words, but I had no way of knowing whether this was a feature likely to occur in 20% adult writers or whether in a representative sample one would find that more people in fact have problems with, for example, past tense forms.

### 3. Spoken, Written, Dictated

The text reproduced as Appendix 2 was presented to the court by the police as a verbatim record of a dictated statement; the accused said some of it was genuinely dictated, some of it was responses to questions and some of it was made up by the police officers concerned. I have reproduced it in full, hoping you will not only take the trouble to read it through but also be stimulated to attempt an analysis yourself afterwards – it is a fascinating text which could play an important role in the development of forensic methodology.

The first sentence is crucial as it is in itself a complete admission of guilt, but also one of the sentences which the accused denied having said:

> (1) I wish to make a further statement explaining my complete involvement in the hijacking of the Ford Escort van from John Smith on Tuesday 28 March 1981 on behalf of the ABC which was later used in the murder of three person in Avon that night.

In arguing that, whatever else it was, it was not a verbatim record of something that had been said, I drew on the work of Halliday on the differences between spoken and written English and in particular on his observations about lexical density:

> On the basis of various samples, I have found that a typical average lexical density for spoken English is between 1.5 and 2 [lexical items per clause], whereas the figure for written English settles down somewhere between 3 and 6, depending on the formality of the writing. (1989:80).

Thus a transcribed spoken sentence conforming to Halliday's 'typical average' would be:

> (4) I drove down to the flats & I saw him up on the roof & I shouted to him & he said that he would be down in a couple of minutes.

which has 8 lexical items spread over 5 clauses, giving a density of 1.6. By contrast, the disputed sentence (1) has at least 25 lexical items, (depending how one counts "28 March 1981", which I chose to regard as three), spread over 3 clauses giving a density of 8.3.

In addition to lexical density this statement contains other features more typical of written text; firstly, to use Halliday again, there is a substantial amount of grammatical metaphorisation which turns processes into nominals. Sentence (1) alone contains "make a statement", "complete involvement", "the hijacking", "the murder".

Secondly, there are several long and complicated theme choices which, in terms of the combination of lexical density and grammatical complexity, demonstrate a linguistic planning which is very unusual in speech:

> (9) It was on the way around to the house at Selly Park that his man who I know to be in the ABC *told me...*

> (13) What happened regarding the hijacking of the van from John Smith that I have already explained in my previous statement *is the truth...*

Thus, this text cannot be a verbatim transcription of spoken English, at least as we have traditionally understood spoken English. However, as the judge pointed out to me, this does not pretend to be a record of typical spoken English; it is a record of *dictated* English. I was forced to agree that we had, as yet, no corpora of dictated texts and that I was therefore extrapolating from my knowledge of the differences between spontaneous speech and written texts and was assuming that the dictated would be in all significant aspects similar to the spoken.

For obvious reasons we are now actively creating a corpus of authentic dictated statements taken from people already accused and awaiting trial or convicted and awaiting appeal. Even so we will still be open to criticism that these are not authentic 'under pressure' statements, but as, so far, no police force has agreed to let us tape or have access to tapes of real statement-taking this is a second best solution. An early observation is that, although the dictated text is more coherent, organised and economical than the spoken, it still seems to conform to Halliday's findings on markedly lower lexical density.

## 4. On the Meaning of Words

As far as I know, John Sinclair's only recent foray into forensic linguistics was an opinion on the ordinary man's understanding of the word *visa*. Apparently in law a visa is not an entry permit but a permit to request "leave to enter". John was asked to provide evidence that this is not the commonly understood use and meaning of the word.

John based his evidence mainly on a 5 million word corpus of *The Times* newspaper, although he supplemented the detailed analysis with reference to the whole of the BCET, totalling, at that time, some 28 million words. *The Times* corpus included 74 instances of "visa" and "visas" in the sense under consideration, of which over 50 collocated with common verbs like "grant", "issue", "refuse", "apply for" and "need". John noted that, although the commonest modifier of "visa(s)" is "exit" it also co-occurs with "entry" and "re-entry":

> you cannot enter an Arab country with an Israeli visa stamped in your passport
>
> British passport holders do not require visas
>
> Non-Commonwealth students who require an entry visa will need a re-entry visa, even if you only leave the country for a couple of days.
>
> (Sinclair 1991)

Thus he concludes that

> the average visitor, encountering everyday English of the type recorded in the corpus, would deduce that a visa was a kind of permit to enter a country...There is nothing...in these examples to suggest that a person who is in possession of a valid visa, or who does not require a visa, will be refused entry. The implication is very strong that a visa either ensures entry, or is not needed for entry. The circumstances of someone requiring "leave to enter" in addition to having correct visa provision does not arise in any of the examples, and the word "leave" does not occurs in proximity to "visa(s)" except in the meaning "depart". (ibid)

This is an example of what can be achieved with a fairly common word and a reasonably small corpus and demonstrates very clearly the usefulness of the method. However, it also shows that it is essential to have a substantial number of instances of the word in question and is therefore in itself a justification for the collection of very large corpora – if for instance one was interested in the word "pimp" which occurs in the existing COBUILD corpora a mere once every 2 million words, one would ideally need a total corpus of some 200 million words.

## 5. Vocabulary choices

The analysis of disputed statements is currently one of the tasks most frequently undertaken by forensic linguists; hopefully it will diminish over time as more and more police forces begin to tape-record all interviews and present a copy of the

tape plus a written summary in evidence. However, there are still many untried cases and appeals, which makes research in this area still important. All statements taken after someone has been accused of a crime and 'cautioned', that is warned that whatever they now say may be taken down and used in evidence against them, are governed by 'judge's rules' which explicitly state that: "The statement should be in the exact words used by the prisoner, it should not be edited or corrected for grammatical errors".

Whereas this is an ideal situation and the rules have been written by people who do not understand the unconscious editing that goes on in and by the most honest of transcribers, it does allow the linguist a foothold to begin to challenge some statements on the basis of unlikely vocabulary choices. You will probably already have noticed, in the Derek Bentley statement, (Appendix 1), the phrases "shelter arrangement" and "brickwork entrance", unlikely expressions for an educationally subnormal illiterate; it will be even more interesting, however, to identify any vocabulary or frequent collocations which are typical of a police register.

In the two statements under consideration above there are several candidate lexical items: *wish, person, involvement, vehicle, hi-jack(ing), (be) approached, (to) name, occur, plain clothes, (in) uniform*; all of these have been noted in authentic statements and reports made by police officers, but only with a substantial corpus will it be possible to say anything about relative frequency, the significance of clustering and the likelihood of an ordinary witness using one or a series of them, in a single statement.

## 6. The Linguistics of Small Corpora

The past decade has given corpus linguists a lot of information about word frequencies and the stability of the phenomena across different types of texts. However, the size of the texts has been typically large: the original COBUILD corpus for example chose samples of 70,000 words. In a forensic context we are very often dealing with extremely short texts, the majority a mere 400 to 700 words long, and unfortunately there is very little research into the frequencies of short texts. No one really knows how small a sample one can reliably work with, at what size significant regularities begin to emerge, whether two samples by the same person can be treated as one larger sample and to what extent regularities vary across registers. This is one of the crucial questions which Coulthard and Winter (in preparation) have begun to investigate.

## 7. On Presenting Evidence in Court

Once forensic linguists have convinced themselves they have something to say they still need to convince the court and this is not an easy task. (It is of course not always easy for the academic linguist to accept that the opposing barrister has no interest whatsoever in the 'truth' of what he has to say.) Courts are, in the main, not knowledgeable about and therefore not particularly interested in statistical significance, and thus one is still very much in the area of rhetorical persuasion. It is

here that concordances could play a significant role. In the second statement presented as Appendix 2, you will have noticed a marked degree of repetition of the item ABC. In court I tried to convince the judge that this was unusual and suspicious, both in terms of redundancy and lack of pronominalisation, but I did so by going through the text with him item by item; in retrospect, I felt that a simple concordance, like the one below would have done the job much more effectively:

| | | |
|---|---|---|
| van from John Smith on Tuesday 28 March 1981 on behalf of the | ABC | which was later used in the murder of three |
| was taken by a person who I do not wish to name but who is a top | ABC | man in the Moseley area wanted to see me & I |
| get a vehicle was to hijack one as it was needed for 7 00pm by the | ABC | & that I was to leave the vehicle in Selly Park |
| the vehicle was required for would be some job connected with the | ABC | but I was not told what the job was. I then went |
| so my mate & me got up & went back into the kitchen where the | ABC | man was & we told him we were going. The |
| I saw on the Ten o'clock News that the van I had hijacked for the | ABC | had been used in the murder of three persons in |
| stocky build with black hair. I am sorry I got involved with the | ABC | in the hijacking of the Ford Escort van which |
| I can only describe the man in the house in Selly Park with the | ABC | man as being small stocky build with black |
| closed over. I heard a conversation from the kitchen and then the | ABC | man called me into the kitchen as he wanted |
| went back to the house in Selly Park I had been at earlier & the | ABC | man let us through the back door. I had parked |
| the kitchen when a car drove up to the back door of the house & the | ABC | man told my mate and I to go into the living |
| was to leave the vehicle in Selly Park and give the keys to the | ABC | man who told me to get it. What happened |
| went into the kitchen and I saw another man there talking to the | ABC | man who told me to give my gloves to the |
| to the house at Selly Park that this man who I know to be in the | ABC | told me that I was to get a vehicle for him as it |

Not only does this concordance emphasise the marked and unnatural repetition of "ABC", it also highlights the marked collocation of the phrase with "man" and also with "murder (of three persons)", "hijack" "vehicle", "job" and "van" – and thus we can see the whole of the disputed charge encapsulated in these collocations, ie that on the instructions of "the ABC man" he "hijacked" a "van" which he knew was going to be used later in a "job", the "murder of three persons".

## Conclusions

In the past few years forensic linguists have had some notable successes, but the future must lie in the creation of a better, standardised and more widely used methodology. What I hope I have shown in the examples above is that any improved methodology must depend, to a large extent, on the setting up and analysing of specialised corpora – only when we know much more about the 'normal' can we be secure in identifying the deviant. One thing, however, is certain: both in planning our corpora and in the subsequent processing of the data we will frequently be drawing on the pioneering research of John Sinclair, whose aim has always been to boldly go where no linguist has gone before.

## References

Coulthard R M and French J P (eds), *Papers in Forensic Linguistics*, (in preparation)

Coulthard R M and Winter E O 'The linguistic analysis of forensic texts', (in preparation)

Fox G (1993)'A comparison of "policespeak" and "normalspeak": a preliminary study', in Sinclair J M, Hoey M and Fox G (eds) *Techniques of Description: Spoken and Written Discourse, A Festschrift for Malcolm Coulthard*, London: Routledge

Halliday M A K (1989) *Spoken and Written Language*, 2nd edition, Oxford: Oxford University Press

Sinclair J McH Unpublished expert opinion on the ordinary man's understanding of the word "visa"

Sinclair J McH (1991) *Corpus, Concordance, Collocation*, Oxford: Oxford University Press

Svartvik J (1968) *The Evans Statements: A Case for Forensic Linguistics*, Göteborg: University of Gothenburg Press

## Appendix 1 Derek Bentley's Statement

I have known Craig since I went to school. We were stopped by our parents going out together, but we still continued going out with each other – I mean we have not gone out together until tonight. I was watching television tonight (2 November 1952) and between 8 p.m. and 9 p.m. Craig called for me. My mother answered the door and I heard her say I was out. I had been out earlier to the pictures and got home just after 7 p.m. A little later Norman Parsley and Frank Fazey called. I did not answer the door or speak to them. My mother told me that they had called and I then ran after them. I walked up the road with them to the paper shop where I saw Craig standing. We all talked together and then Norman Parsley and Frank Fazey left. Chris Craig and I then caught a bus to Croydon. We got off at West Croydon and then walked down the road where the toilets are – I think it is Tamworth Road.

When we came the place where you found me, Chris looked in the window. There was a little iron gate at the side. Chris then jumped over and I followed. Chris then climbed up the drainpipe to the roof and I followed. Up to then Chris had not said anything. We both got out on to the flat roof at the top. Then someone in a garden on the opposite side shone a torch up towards us. Chris said: 'It's a copper, hide behind here.' We hid behind a shelter arrangement on the roof. We were there waiting for about ten minutes. I did not know he was going to use the gun. A plain clothes man climbed up the drainpipe and on to the roof. The man said: 'I am a police officer – the place is surrounded.' He caught hold of me and as we walked away Chris fired. There was nobody else there at the time. The policeman and I then went round a corner by a door. A little later the door opened and a policeman in uniform came out. Chris fired again then and this policeman fell down. I could see he was hurt as a lot of blood came from his forehead just above his nose. The policeman dragged him round the corner behind the brickwork entrance to the door. I remember I shouted something but I forget what it was. I could not see Chris when I shouted to him – he was behind a wall. I heard some more policemen behind the door and the policeman with me said: 'I don't think he has many more bullets left'. Chris shouted 'Oh yes I have' and he fired again. I think I heard him fire three times altogether. The policeman then pushed me down the stairs and I did not see any more. I knew we were going to break into the place. I did not know what we were going to get – just anything that was going. I did not have a gun and I did not know Chris had one until he shot. I now know that the policeman in uniform is dead. I should have mentioned that after the plain clothes policeman got up the drainpipe and arrested me, another policeman in uniform followed and I heard someone call him 'Mac'. He was with us when the other policeman was killed.

## Appendix 2 Disputed Statement

(1) I wish to make a further statement explaining my complete involvement in the hijacking of the Ford Escort van from John Smith on Tuesday 28 March 1981 on

behalf of the ABC, which was later used in the murder of three person in Avon that night. (2) The truth about what happened from the start is that I wasn't approached in the Ladbroke as I had said before but I was taken by a person that who I do not wish to name but who is a top ABC man in the Moseley area wanted to see me & I was to go to a house in Selly Park Moseley where he would meet me. (3) I went to this house in Selly Park & a woman there told me that the man I was looking for was down at the flats in Goodby Gardens. (4) I drove down to the flats & I saw him up on the roof & I shouted to him & he said that he would be down in a couple of minutes. (5) I forgot to say that this would have been a short time after 4.00pm on Tuesday 28 March 1981. (6) I waited a few minutes & the man I was to meet came down & got into the car & I drove him around to the house in Selly Park. (7) We both went into the house & I saw his girlfriend Nicola painting in the living room. (8) I only stayed a few minutes & left. (9) It was on the way around to the house at Selly Park that this man who I know to be in the ABC told me that I was to get a vehicle for him as it was needed for a job that night. (10) I understood that the job that the vehicle was required for would be some job connected with the ABC but I was not told what the job was. (11) I then went from Selly Park to my mates house & I told him that I needed his help to get a vehicle & I would call & pick him up around 6.30pm. (12) I explained to him that the only way we were going to get a vehicle was to hijack one as it was needed for 7.00pm by the ABC & that I was to leave the vehicle in Selly Park & give the keys to the ABC man who told me to get it. (13) What happened regarding the hijacking of the van from John Smith that I have already explained in my previous statement is the truth up until we took the van to Selly Park. (14) Myself & my mate went back to the house in Selly Park that I had been at earlier & the ABC man let us in through the back door. (15) I had parked the van in the position that I have already shown on the map. (16) The three of us were in the kitchen when a car drove up to the back of the house & the ABC man told my mate & me to go into the living room. (17) We sat down in the living room & I remember that it must have been after 7.00pm as 'Top of the Pops' was on the T.V. (18) I heard somebody come in through the back door of the house into the kitchen, I couldn't see as the door between the kitchen & the living room was closed over. (19) I heard conversation from the kitchen & then the ABC man called me into the kitchen as he wanted the gloves I had. (20) I went into the kitchen & I saw another man there talking to the ABC man who told me to give my gloves to the other man which I did. (21) The gloves were black woollen gloves which I owned for ages. (22) I then went back into the living room & I heard the back door close so my mate & me got up & went back into the kitchen where the ABC man was & we told him we were going. (23) The other man had left the house & when we went outside the van we had hijacked had gone so we got into my car & I dropped my mate home & I went on to work at the Chinese Take-Away. (24) The only other thing that occurred was that I delivered a Take Away meal from the Chinese to that same house in Selly Park later that night. (25) I'm not sure what time. (26) It was later that I heard about the murders in Avon & I saw on the Ten o'clock News that the van I had hijacked for the ABC had been used in the murder of three persons in a mobile shop. (27) I can only describe the man in the house in Selly Park with the ABC man as being small, stocky build with black hair. (28) I am sorry I got involved with the ABC in the hijacking of the Ford Escort Van which was used in the murder of three persons in Avon that night.

(29) I have read the above statement and I have been told that I can correct alter (sic) or add anything I wish. (30) This statement is true I have made it of my own free will.

# THE PRAGMATICS OF TEXT AVERRAL AND ATTRIBUTION IN ACADEMIC TEXTS

*Angele Tadros*

## 1. An outline of the paper

This paper examines interaction in written text through the interplay between the notions of text averral and attribution (Sinclair, 1988). Text averral is evidenced in the unmarked parts of the text, where the utterances are assumed to be attributed to the author. Attribution, the counterpart of text averral, is the marked case where the sources of authority are clearly signalled. The focus of this paper is on one pragmatic aspect of the interaction between writer and reader. The data are drawn from texts in linguistics, sociolinguistics and discourse analysis. The stated purposes of the three texts are different and there is also variation in reader orientation. It is the purpose of this paper to discuss whether differences in author purpose and reader orientation are reflected in the realizations of the pragmatics of text averral and attribution. It is hoped that this study will add to our knowledge about the characteristics of different types of text, and illuminate the way for students who find themselves lost amidst the echoes of the multiple voices they hear within the same text.

## 2. Introduction

At one time pragmatics was regarded as the 'wastepaper basket' of linguistics. Whatever could not be adequately handled under linguistics was dumped in the wastepaper basket. Recently, this attitude has changed, and pragmatics is now established as an autonomous field of study (Levinson, 1983).

A number of topics in pragmatics have received due attention: speech acts (Austin, 1962; Searle, 1975), conversational implicature (Grice, 1981), politeness strategies (Myers, 1989), conversation analysis (Sacks et al, 1974), to mention but a few. However, although the pragmatic notion of attribution and, more specifically, citation has received a great deal of attention (see Swales, 1981, 1990; Cronin, 1981; Francis, 1986; Harvey, 1992), little has been written about the interplay between attribution and its counterpart text averral, apart from the important discussions of related issues in Sinclair (1983, 1985, 1986 and 1988).

This paper attempts to investigate the notions of text averral and attribution in a corpus drawn from the areas of linguistics, sociolinguistics and discourse analysis,

to find out whether the stated purposes of the authors of the texts are reflected in the interplay between those notions. First, the corpus will be described, followed by a discussion of the notions of text averral and attribution as evidenced in the corpus. In the third part the two parameters of author purpose and reader orientation will be discussed in relation to those notions. Lastly, there will be a conclusion in which theoretical and pedagogical observations will be made.

## 3. Corpus

The corpus for this study consists of texts drawn from three closely related disciplines: linguistics, sociolinguistics and discourse analysis. The texts are:

1. *An Introduction to Linguistics* by L. B. Crane, E. Yeager and R. L. Whitman, pp 102-116, published by Little, Brown & Co. in 1981;
2. *Sociolinguistics* by Roger Bell, pp 89-115, published by Batsford in 1976;
3. *Towards an Analysis of Discourse* by John Sinclair and Malcolm Coulthard, pp 1-18, published by Oxford University Press in 1975.

The texts were examined from the perspective of text averral and attribution to find out how much is revealed about the purposes of the authors and intended readers, and whether what is revealed corresponds to the authors' purposes and intended readers as explicitly stated in the prefaces and introductions[1].

Before we proceed with the discussion, a brief word on writer purpose and reader orientation is in order.

### 3.1. *An Introduction to Linguistics* (C et al)

The authors of this text announce that their book has been designed as 'an introduction to linguistics for undergraduates and beginning graduate students' (v-vi), though other interested readers may find it useful. From the start, we are made aware of the conceptual distance between the authors and their readers. Because the readers are assumed to be new to the field, they are given a great deal of guidance. The authors orientate the readers to the approach they are adopting i.e. that of transformational-generative grammar, since 'they have found this a useful way to organize and present the material in the text'; also it has been the dominant approach in 'the past twenty five years'. This, of course, is not to be questioned, since their naive readers are not expected to doubt the rationale behind the choice. However, honest to their mission, the authors confess that there are many opposing theories (p. vii) which the students will surely encounter at a later stage. But, for now, they are advised to keep an open mind and to question those statements which may appear somewhat dogmatic (p.vii).

Of course, as any experienced teacher would know, such admonitions are easier said than followed, and as will be shown later the authors work hard at marching their inexperienced readers into an arena where opposing viewpoints and theories are annihilated and only the banner of transformational-generative grammar is hoisted.

### 3.2. *Sociolinguistics* (B)

In the introduction, the author gives the caveat 'This book is by no means an introductory textbook' (p. 11). The book aims to integrate materials and approaches from

various sources in order to give the student an integrated picture of the whole field and the teacher a very wide range from which to select his materials for teaching.

3.3. *Towards an Analysis of Discourse* (S&C)

In this text we have a different situation from that of the other two -- a situation in which the authors offer themselves as researchers. 'This monograph describes the findings of a research project, ...'. This orientation is revealed in various ways throughout the text – in the introduction, in the chapter on the review of the literature, in the fact that there is real data presented for analysis and in the detailed critical bibliography towards the end of the book. It would have been unexpected had we encountered these features in the first two texts.

The book is distinctive in that specific chapters are oriented towards specified audiences – chapter three, for instance, caters for both the general reader who will be 'interested in the broad outlines of the system' (p.19) and the researchers who 'will want to use the system in their own research work' (p.19), whereas chapter four, which includes meticulously analysed texts, caters for the reader who needs a lengthy exemplification in a new area where there is no established methodology. Chapters one and two, which constitute our corpus, aim to place the current research in the context of previous research literature and to move towards the creation of a new model which will be suitable for the analysis of the data.

## 4. Text Averral and Attribution

Text averral and attribution are basic notions for the organization of interaction in written text. The assumption is made that the author of a non-fictional artefact (Sinclair, 1986) avers every statement in his or her text so long as he/she does not attribute these statements to another source – whether that source is other or self. Averral is manifested in various ways in the text – negatively, through absence of attribution, and positively, through commenting, evaluating or metastructuring of the discourse. Attribution, on the other hand, is signalled in the text by a number of devices of which reporting is an obvious one.

To illustrate the notions of text averral and attribution, let us look at the following examples from the data:

*Example 1*

> There are of course techniques one can use in conversation to make it more likely that the discourse will continue on the lines you intend. (S&C p.5)

*Example 2*

> The traditional definition of a sentence as 'a complete thought' goes back to the ancient Indian and Greek Grammarians, ... (C et al p. 102)

*Example 3*

> but the definition lacks persuasiveness. (C et al p.102)

Examples 1 and 3 illustrate the notion of averral. In example 1 the authors aver that there are techniques... and we assume that what they aver corresponds with

what they believe. This averral is negatively manifested through absence of attribution. If asked, 'what are those techniques...?' they cannot answer 'There are none'. In example 3, also an instance of text averral, the writer surfaces as evaluator. The 'but' to use Sinclair's terminology (Sinclair, 1985) indicates a change of posture, i.e. in this case a return to averral; as a result the authorial voice is heard. This, however, does not mean that evaluation cannot be attributed to someone other than the author as is clear from the following made-up example:

*Example 4*

> ... but many linguists regard such techniques as unreliable.

Incidentally, it will be noted that the 'but' is still the author's which indicates a return to text averral, although the proposition is attributed to others. In other words, the author is saying 'But I aver that many linguists...'. The status of the evaluation remains unaltered.

Concerning example 2, we find that there is attribution. This surfaces most obviously in the quoted piece of text, which is attributed to the ancient Indian and Greek grammarians, but less obviously in the use of the modifier 'traditional' which probably the authors would not choose to align their approach with.

## 5. Text Averral and its realizations in the corpus

In one sense text averral refers to the unmarked part of the text – the part where there are no attributions; in another sense it refers to all those parts where there are positive indications of the author's presence.

We noted earlier that text averral is negatively indicated through the absence of attributions where an underlying 'I aver to you that' can be postulated. The verb 'aver' is the signal of the default condition. The default hypothesis states 'that in the absence of contra-indications, a linguistic item has the same function as its predecessor' (Sinclair, 1985:10).

*Example 5*

> Syntax is the way words are put together to form phrases and sentences. (C et al p.102)

*Example 6*

> A closed system can be related to its models by means of categorical rules, which, if correctly formulated, always apply. (B p.91)

*Example 7*

> Much of everyday language use is not designed to be verbally explicit, direct and literal, but can achieve its ends in subtle ways... (S&C p.12)

It will be observed that the three pieces of text above are not attributed by their respective authors to any source and so are instances of negative averral. It is assumed that the authors believe in what they are saying, at least in terms of the discoursal world they have established. This world continues to dominate the discourse until attribution is signalled. If no attribution is signalled, the default condition obtains. Positively, text averral is manifested via the use of first person

pronouns. Since two of the texts under study (S&C and C et al) have joint authorships there is no way in which authorial presence will be realized by 'I', whereas the third text is written by one author and both the singular and plural pronouns are possible.

Not all cases of 'we' have their referents as joint authors. It is interesting to note that in examples 8 and 9 below, taken from the same corpus, the 'we' in example 8 is interpreted as referring to both authors and readers, whereas the 'we' in 9 refers to the authors unaccompanied by readers.

*Example 8*

> To take a concrete example, if someone has a pile of objects in front of him and says
>
> This is a wonk, this is a dibble...
>
> we do not know what is happening. (S&C pp. 15,16)

*Example 9*

> We felt the need to begin with a form of discourse which had more structure and direction. (S&C p.5)

In example 8 above , 'we' includes 'you readers' The reason for this, I suggest, is that the piece of text is marked for exemplificatory purposes by "To take a concrete example..." and examples are given by authors for the benefit of readers in an interaction, so the readers are included. In contrast, in example 9 'we' refers to the authors alone, since the piece of text includes an act which is part of the procedural steps taken by the authors while they were conducting their research. There is no way in which readers could be participants in such an undertaking.

Other examples where text averral is realized by the inclusive 'we' are:

*Example 10*

> Every day we utter and understand sentences that we have never heard before. (C et al p.106)

*Example 11*

> We can therefore think in terms of primary relationships within primary social groups or, to use a term frequent in sociolinguistic description, domains. (B p.103)

Often the inclusive 'we' is used when the author is revealing the structure of the discourse to the reader by means of metastructuring devices such as Recapitulation or Advance Labelling (Tadros 1981,1985). In Recapitulation the author tells the reader what he/she has already done in the text, whereas in Advance Labelling he/she tells the reader what he/she is going to do. These devices bear some resemblance to Sinclair and Coulthard's focusing move (S&C,1975). They are illustrated in examples 12 and 13.

*Example 12*

> In the previous chapter, we listed what appear to be the major components and functions of a model of language in use. We did not, however attempt... ( B p.89)

The above example illustrates Recapitulation, an instance of text averral. The author refers to an act he has already performed, i.e. 'listed', but by using 'we' he is including the readers as participants in the act.

*Example 13*

> In this chapter, we examine first the problems with the traditional and structuralist approaches to syntax and then... ( C et al p.102)

In the piece of text above, the authors label in advance a discourse act 'examine'. This helps focus the readers' attention on what they want to do. It will be observed that the use of the first person plural helps to achieve solidarity (Myers, 1989) and gives the reader a sense of assurance that he/she is not being left out of what is happening in the discourse.

It is interesting to note that all acts associated with the first person pronouns indicate intellectual activities. Bazerman (1981) p. 367 summarizes such activities thus: '... statement making..., making assumptions..., criticizing statements, and placing knowledge claims within other intellectual frameworks... .' Apart from the use of first person pronouns as possible realizations of text averral, there is also the use of comments and evaluations, unless of course these are attributed to a source. Usually there are lexical signals pointing to the evaluation or comment, or else one of the items which indicate change of posture (*but, however, in fact, perhaps*, etc.).

*Example 14*

> This was an exhausting chore of no great merit for in the end the patterns expressed nothing more than what was already known: that a variety of sentence types existed. (C et al p. 105)

*Example 15*

> But with this proviso in mind, we shall continue to make use of the distinction, because of its considerable value as a marker of linguistic choice – we are, after all, still attempting to correlate linguistic with social structure, the basic goal of sociolinguistics. ( B p.103)

*Example 16*

> Two major drawbacks of using such performance data are firstly and obviously that a child who does not use a particular structure... Secondly and more seriously, ... ( S&C p.3)

All the above pieces of text are averred by their authors. In example 14, the authors are negatively evaluating some activity mentioned in a preceding piece of text. The negative evaluation is clearly signalled both lexically and grammatically: 'an exhausting chore', 'of no great merit'. and 'nothing more than what was already known'.

In example 15, there is positive albeit conditional evaluation. Positive evaluation is manifested lexically in ' of considerable value', and also in the implicated meaning deriving from the fact that there is an expressed desire on the part of the author to correlate things rather than to separate. The evaluation is averred and this averral is coupled by the use of the first person pronoun 'we'.

Lastly, in example 16, the authors signal their negative evaluation lexically by means of the item 'drawbacks'. Here again, the evaluation is averred.

Before moving to the next section, which is concerned with attribution, let us summarize the position pertaining to the various realizations of text averral. Text averral is realized in the corpus in the following ways:

1. negatively – by absence of attribution
2. positively – through the use of:
   a. first person pronouns
   b. comments and evaluations

2a. and 2b. are subject to certain conditions, as discussed above.

## 6. Attribution and its Realizations in the corpus

Attribution is the marked part of the text, marked as belonging to another source i.e other than that of the text being created at the moment of the utterance. It is the counterpart of text averral and the interplay between these two notions is an important aspect of interaction in written text.

According to Sinclair (1988) attributions are reports in the text which have the effect of transferring responsibility for what is being said. Attributions in the corpus under study have been realized by a variety of linguistic resources. These may be classified into two broad categories on the basis of whether they are accompanied by citations or not:

### 6.1. *Group A: Citational*

This group includes what Swales (1990) calls integral and non-integral citation. Thus: 'An integral citation is one in which the name of the researcher occurs in the actual citing sentence as some sentence-element; in a non-integral citation, the researcher occurs either in parenthesis or is referred to elsewhere by a superscript number or via some other device'. (Swales, 1990: 148). It is worth noting that in discussing citational attribution, C *et al* will not appear. This is because, with only one exception, no citations of this type have occurred in the corpus drawn from their text. This point will be taken up later in this paper. Integral and non-integral citations are illustrated below:

*Example 17*

> Transformational grammar by its very nature is unsuited to handling such context-dependent meanings, although Katz and Postal (1964) do discuss the case of the sentence 'You will go ...' (S&C p.11)

*Example 18*

> Laboratory exercises, designed to test the relationship between group effectiveness and communication net (Leavitt, 1951), demonstrate several configurations, some of which ... (B p.107)

In example 17, citation is an integral part of the citing sentence. The nouns 'Katz and Postal' function as the subject element in the subordinate clause structure. In example 18, on the other hand, the cited noun is imprisoned in parenthesis and so

does not come to life as an element in the structure of the citing sentence. Both integral and non-integral citations may be accompanied by direct quotations.

6.1.1. *Integral citations with direct quotations*

*Example 19*

> It follows, then, that rather more of the rules of sociolinguistics turn out, given as Searle (1969) consistently puts it (p.126) 'normal input and output conditions obtain', to be categorical in nature and hence... (B p.101)

In the above piece of text, the part in quotations is attributed to Searle, the noun 'Searle' functioning as subject in the subordinate structure. The rest of the text is averred as is shown by the signals indicating change of posture: 'It follows, then' and also 'hence'. These items indicate that the author is making deductions and thus averring the text.

*Example 20*

> Firth (1935) had observed that in conversation 'we shall find the key to a better understanding of what language really is, and how it works.' (S&C p.3)

In this example there is a quotation in inverted commas and this is attributed to Firth. The citation is integral, since the noun 'Firth' functions as subject in the sentence.

6.1.2. *Integral citations without direct quotations*

*Example 21*

> This model is based on that suggested by Ervin-Tripp (ibid) but contains a few minor adjustments. (B p.96)

*Example 22*

> Only the theoretical papers of Hymes (1962,1964,1967,1972) bore exactly on what we were attempting. ( S&C p.10)

In the above examples (21 & 22) the citations are integral: the nominal 'Ervin-Tripp' functions as agent in sentence structure and 'Hymes' as object to the preposition 'of'. In neither case is there a direct quotation.

We now leave integral citation behind and move on to the other type of citation – the non-integral, where the name of the individual whose work is being cited does not function as an element in the sentence structure. As noted earlier, this type can also occur with or without direct quotations.

6.1.3. *Non-integral with direct quotations*

*Example 23*

> One statement of the aims of sociolinguistics (Fishman, 1970, p.3) emphasizes the importance of the discovery and the specification of sociolinguistic rules in a very clear way '... sociolinguistics seeks to

> discover the societal rules or norms that explain language behaviour and the behaviours toward language in speech communities'. (B p.92)

*Example 24*

> Secondly and more importantly, despite the efforts of the 'sentence stretchers' (Boesen 1966), there is no principled reason why grammatical complexity should be 'better'. (S&C p.3)

In examples 23 and 24 the individuals whose works are being cited are placed in parentheses and are not allowed by the creators of the texts to take part in the arguments presented. In example 23, the statement attributed to the source is quoted, probably because an attempt at paraphrasing might not have kept the intended meaning. In example 24, the authors quote from the source the term 'sentence stretchers' and the item 'better' both of which cannot be interfered with.

*6.1.4 Non-integral without direct quotations*

*Example 25*

> Indeed, it is just this learning of how to choose, which lies at the root of the socialization process in the child (Fichter, 1971 p.218) and, ... (B p.105)

*Example 26*

> Both tests and measurements were crude but show the beginning of the current emphasis on communicative, and not simply linguistic, competence (Hymes, 1972, Candlin, 1972,Widdowson, 1971). (S & C p.3)

It will be noticed that in examples 25 and 26 there are no direct quotations but there is an indication that certain ideas are attributed to the sources in parentheses. Text averral in example 25, as in the use of 'indeed' and the anaphoric use of 'this', merges with attribution without rendering the text ambiguous. To help us arrive at what is being attributed we could try to substitute an integral citation as in: 'It was pointed out by Fichter (1971 p. 218) that learning how to choose lies at the root of the socialization process in the child'. But such a transformation would have necessitated changes in the organization of the text.

Similarly in example 26, the part which is attributed to the sources in parenthesis is not the evaluative comment about the tests. "Both tests and measurements were crude" is an instance of averral, but the signal 'but' indicates a change of posture – a forthcoming attribution – which we find in the ideas implicated by the terms 'communicative' versus 'linguistic'. These ideas are attributed to the sources in parenthesis.

In the same way that authors cite others, they can also cite themselves. This phenomenon – self citation – has occurred in the corpus from Sinclair and Coulthard, presumably because they are both researchers and writers. The text is written jointly, but there are citations to individual previous works. In such cases, the author is referred to in the third person just as would be the case with any other source. The motives for self citation in the text under study might relieve Swales

of some of his woes as regards motivations for self citation (Swales 1990, p.6). The authors in our text are self citing in order to reveal the inadequacies of their earlier approaches.

*Example 27*

> The approach proposed by Sinclair (1966) was diametrically opposite to Chomsky's. He suggested examining real examples with all their performance features. (S&C p.2)

*Example 28*

> To approach the assessment and measurement of disadvantage from a different angle Wight and Norris (1970) and Coulthard (1970) devised a series of Language Function Tests. (S&C p.3)

In examples 27 and 28 there are integral self citations. In 27, Sinclair's work is cited, the purpose being to show the shortcomings of this earlier approach, together with the following work of 1968. These shortcomings are revealed in the following piece of text: 'The major drawbacks of using such performance data...' (S&C p3). The piece of text in example 28 is attributed to Coulthard among others. It is also followed by a comment, an instance of text averral as in 'Both tests and measurements were crude...' (S&C p.3).

6.2. *Group B: Non-Citational*

The second category of attribution is the non-citational. Here no actual source is cited but there are attributions to 'disciplines', 'schools of thought, 'groups of researchers' and the like. This is the type most prevalent in textbooks and is in fact, with one exception, the only type occurring in the corpus from C et al. The reasons for this phenomenon will be discussed in the part dealing with author purpose and reader orientation.

*Example 29*

> When the transformationalists succeeded the structuralists they also made the sentence central to syntax;... defined parts of speech..., they said a sentence..., sentences were recognized to be sentences intuitively. ( C et al p.103)

*Example 30*

> Linguistics, in this century, has attempted to take a descriptive rather than a prescriptive orientation to its data and sought to build up a body of rules,... (B p.90)

*Example 31*

> Philosophy, like linguistics, has concentrated its attempts on referential uses of language. (S&C p.13)

In the three examples just given, there are attributions without citations. In example 29, there is a series of attributions to the transformationalists signalled by 'made', 'defined', 'said', 'were recognized', all referring to cognitive processes. In example 30, reference is made to the whole discipline of linguistics as regards how

it approached its data: 'has attempted to' and 'sought'. Again no citations are made. In the last example (i.e. 31) there is reference to the discipline of philosophy to which is attributed the concentration on referential uses of language. The fact that no citations are made suggests that no authority or substantiation are deemed necessary to support the attributions made by the authors. The readers will accept what the authors are saying without proof.

## 7. Author Purpose and Reader Orientation

### 7.1. *Crane et al*

As pointed out in section 3, *An Introduction to Linguistics* is an introductory textbook and hence there will be a conceptual distance between authors and readers. With one exception, attributions in the corpus are unaccompanied by citations. We assume the authors of the text feel they do not need to establish the credentials of their work; they do not need to cite an authority to support an opinion or a point of view. Are they not themselves an authority addressing a body of readers new to the discipline? Citation will shake their position as possessors of the key to knowledge. There is an analogy between this situation and that of the relationship between a mother and her child. If a mother, instead of telling her child what to do, keeps saying "Dad says 'do this'". Dad says 'do that'", the child will lose his respect for his mother who has no authority since she is constantly 'citing' father. The same is true of the relationship between the author of an introductory textbook and the reader. The observation about the absence of citations finds support in Myers (1992) as cited in Swales (1991, in mimeo) who says that in Myers' recent discussion of textbooks in the sciences he 'instances the absence of hedging and the paucity of references to the primary literature'.

In the corpus drawn from C et al, there is a considerable number of attributions without citation. The authors attribute theories, approaches, opinions, definitions to others in order to knock them down so that they can set up the approach they align themselves with.

In their preface and notes to the student reader, the authors make their intentions explicit. They declare they are going to approach syntax 'primarily from the transformational generative point of view, with introductory sections on previous approaches' (p.vi). Previous approaches or "opposing viewpoints and theories" (p.vi) are indeed discussed, but with the intention of exposing their inadequacies or negatively evaluating them and leading the students along the path of transformational-generative syntax.

An example of how this is achieved through the pragmatics of text averral and attribution is given below:

*Example 32*

> The definition of a sentence became a real problem for the structural linguists; for they ... discovered... . They rejected the easiest definition – ... because they believed ... . Eventually they adopted... (C et al p.103)

In example 32, the attribution is reiterated by a series of cognitive verbs. First, there was a 'real problem for the structural linguists'; then, in a series of clauses

come the reiterated attributions: 'they discovered', 'they rejected', 'they believed'. 'they adopted' – all verbs of the non-factive or semi-factive type which do not commit the writer to the truth of the proposition expressed. These are then followed by the following evaluation, an instance of text averral:

> But everyone knew that this definition was terribly weak, just as they knew that the likelihood of a better definition was remote. (C et al p.103)

Here the authors conceal themselves behind the word 'everyone'. This is followed by the factive verb 'knew' ,which unlike the verb 'thought' cannot be rebutted even when it is in the past tense, as the following will show:

> Everyone knew that this definition was terribly weak but it wasn't.

contrasted with

> Everyone thought this definition was terribly weak but it wasn't.

Thus the propositions after the factive verb 'knew' have to be accepted as facts, and although Winter (1977) and Hoey's (1983) interlocutors would have produced the inevitable questions:'In what way was that definition weak?' and 'why was the likelihood of a better one remote?', our text creators avoid the company of such interlocutors. They have nothing to do with inquisitive companions – 'You reader who are assumed to be a child in the discipline should not ask such questions', reminiscent of a curious child who wants to know things beyond his cognitive level to whom his mother produces the admonition, 'When you grow up you will find answers to such questions', or else gives a part-of-the-truth answer.

With the rivals removed, the way is clear for the transformational-generative approach, which has been pre-selected. Here evaluations generally tend to be positive – the approach is for example, 'much simpler', and if there are negative evaluations they are mitigated. Thus at first we are given the following negative evaluation 'no definition is necessary', 'sentences were recognized to be sentences intuitively'. 'Although this definition inspired research and insight, it certainly has its shortcomings' (C et al p.103). However it immediately turns out that the 'shortcomings' are not serious – since they concern matters of judgement: 'speakers' intuitions often disagree'.

So we can see the extent to which readers are led by the authors of the textbook.

The only instance of citation that occurs in the corpus (p.104) is made in order to expose the ludicrous examples and the flawed rules of prescriptivism. It is unfortunately too lengthy to quote here.

In this part I have attempted to show how the authors' purpose and reader orientation are reflected in the careful manipulation of text averral and attribution.

### 7.2. *Bell*

Moving from Crane et al to Bell, we hear a different story. Bell's anticipated readers are novices in sociolinguistics, but should come armed with techniques adopted in general linguistics. In this sense the 'book is by no means an introductory textbook' (p.11). It aims to provide an integrated picture of the whole field of sociolinguistics.

Another group of audience is the teachers – who are warned that the book 'probably does not recommend itself to a teacher in search of a textbook' (p.11).

In what follows, I will try to show how the author's stated purpose and reader orientation are manifested in the pragmatics of text averral and attribution.

In attempting to relate purpose to audience, I will start with the teachers since they can be dismissed before their students. The author presents the teacher with a wide variety of approaches to choose from – all manifested naturally enough by means of attribution, and more specifically via citation of what Swales (1990) calls the integral type. Examples of attribution structures are the following, which come in a sequence:

*Example 33*

> ...Fishman (1970) which 'attempts to ...' Burling (1970) approaches the subject as an anthropologist ... Pride (1971) writes from the view point of a linguist... Robinson (1972) writes as a social psychologist... And, finally Trudgill (1974) like Pride, comes to the subject as a linguist... (B p.12-13)

After having presented all these approaches, the author then goes on to place his book within the context of the field. Text averral is clearly marked by reference to the author's own text:

*Example 34*

> This present book probably falls, in its orientation, between those of Fishman and Trudgill since it attempts to cover both the sociology of language and sociolinguistics. It differs... (B p.13)

The writer's purpose in referring to all the scholars mentioned in example 33 is given clearly in his concluding statement in the introduction: '...that the individual teacher has available to him a very wide range of easily accessible texts' to choose from (p.13).

Having thus given our teachers an early dismissal, we move on to a discussion of how the author attempts to achieve his aim of 'integrating materials and approaches from diverse sources' (B p.12) and how this is revealed in the pragmatics of text averral and attribution.

Unlike the textbook authors Crane et al, the author here acts as a pacifist whose aim is not to smash but to put things together and as a result create a 'cumulative' and 'integrated approach' to the study of sociolinguistics. So it is to be expected that such aspects of text averral as evaluation and comments will be mitigated, and attributions will have a variety of functions – to provide support, to summarize an opinion or point of view, to give credentials and so on.

The author avers that one definition of a rule might be along the lines of 'a formal statement...' and then comments, 'such a view would be consistent with the usage of many philosophers of science (e.g. Nagel 1961 pp. 90f.)'. No quarrels or confrontations. There is an orientation towards integration – even when he discusses two contrasting types of rule – descriptive and prescriptive – and so we find an evaluation which reflects the policy of integration:

*Example 35*

> Indeed, where the sociolinguist is involved in the application of his discipline ..., he will find himself needing to adopt a position of 'enlightened prescriptivism' based on evidence...

> However, just as..., so the sociolinguist is essentially concerned to specify the descriptive rules... (B p.90)

Thus for the author, it is not a question of *either/or*, but a question of an aspect of this and an aspect of that. But before reaching this position, he has attributed prescriptive rules to traditional grammar and rhetoric (p.90) and descriptive rules to 'linguists in this century'. Both attributions are followed by evaluations, not of the destructive type encountered in the textbook, but of the mitigated type.

*Example 36*

> Often such prescriptive rules have their origin in Greek or Latin convention, and more often than not, bear little relation to actual present-day English usage (see Palmer, 1971 pp.13-26). (B p.90)

The mitigation is manifested in the shifting of the responsibility from traditional grammar and rhetoric to Greek and Latin convention and, further, the author directs the reader to Palmer for confirmation.

The other instance of evaluation in this sequence is that concerning descriptive rules where the criticism is couched in a subordinate expression (italicised below) and is accompanied by mitigating items such as 'in fact' and 'somewhat' as in :

*Example 37*

> ... and sought to build a body of rules based, ... on empirical evidence derived from the observation of actual, *though in fact somewhat idealized* speech, (B p.90) (my italics)

Throughout the corpus, there is a constant interplay between text averral and attribution related to the author's purpose and reader orientation, but space would forbid the discussion of other instances of this phenomenon.

### 7.3. *Sinclair and Coulthard*

The text by Sinclair and Coulthard is different from the other two texts in that in this text the authors themselves are researchers. They have already done their research into the English used by teachers and pupils and submitted a report. In the text under study, they are presenting an updated version of the report in the form of a book. The anticipated readers range from general readers to specialized researchers who might be interested in adopting the systems of analysis. So, in order to cater for these groups of readers the various chapters are written with different orientations to readers. The parts of the text selected for this study are chapters 1 and 2 – the introduction and 'Short Review of the Literature'. In their review of the literature, the authors dispose of previous models because they are not suitable for their data. None of the approaches or schools of thought attributed is without inadequacies. Even the authors' previous attempts (Sinclair, 1966, 1968) and (Coulthard, 1970) (p.3) are lacking in some respect or another. This is compatible with their ultimate purpose of designing a discourse model suitable for their data.

Their area involves a critique of predecessors – renowned people like Chomsky, Firth, Halliday, Hymes and themselves. They will have to approach the problem of reviewing the literature tactfully so as not to hurt scholars who are devout adherents to this school or that.

Thus, if researchers want to analyse discourse, they should not go to Chomsky's transformational model of the mid-sixties, nor Katz and Fodor (1963) nor Sinclair (1966) (p.2) nor Sinclair (1968) nor Wright and Norris (1970) nor Coulthard (1970) nor Firth (1965) (all on p.3) nor Halliday (1967) nor Hasan (1968) (p.8) etc.

The readers are made aware that there is no intention to undermine the value of previous work by being constantly reminded of the authors' purposes: 'Our interests again were in the function of utterances...' (pp. 3-4); 'this was the very thing we wanted to avoid...' (p.10); '...but he was not working within a linguistic framework...' (p.17); 'however the level of language function in which we were centrally interested, is neither ... nor... . It is rather the level of the function of a particular utterance ... ' (p.13). Previous models are inadequate – not in an absolute sense – but relative to the authors' purposes. The readers become ultimately convinced that no model in existence is able to handle the data. There is throughout an interplay between text averral and attribution, as each model is brought for inspection (attribution) and then dismissed (averral via evaluation).

To illustrate the interplay between attribution and averral, let us consider the following example:

*Example 38*

> The only attempt within sociolinguistics to describe the structure of a spoken text is by Mitchell (1957) who describes the language of buying and selling in Cyrenaica. He divides the transactions into a series of stages and then discusses... (S&C pp. 9-10)

*Example 39*

> However, the stages are not isolated on linguistic criteria, there are no boundary markers ... the numbering of stages does not ... imply ... . In fact, stages are simply defined by the kind of activity... . Neither does he attempt to provide a linguistic structure... but simply ... This was the thing we wanted to avoid... . Our interest was always in the linguistic structure of discourse. (S&C p.10)

In example 38 a model is attributed to Mitchell via integral citation and the use of verbs 'describes', 'divides', 'discusses' all related to cognitive processes. The model is rejected as is evidenced in example 39, not because it was flawed, or inconsistent or obscure, but because the model's interests were incompatible with the authors' interests. They were looking for a linguistic structure for discourse and the model in question used non-linguistic criteria.

The move from attribution is signalled by 'However'; then follows a series of negative items: 'not isolated', 'no boundary markers', 'does not ... imply', 'In fact...simply', 'neither does he...' 'but simply', after which there is a comment indicating the inadequacy of such a model for the authors' purposes: 'This was the very thing we wanted to avoid'. The passage culminates in a summary statement of why the attributed model would not do for them: 'Our interest was always in the linguistic structure of discourse'. The use of 'always' in this summary statement seems to have some interactive significance. It has the effect – or so it seems – of bringing to the attention of the readers the fact that the declared interest in the linguistic structure of discourse is not a momentary justification for rejecting the

model – but that even before they examined the model their interest was always in the linguistic structure of discourse.

There are many more examples in the data – too numerous to be included in this discussion – of the way the interplay between attribution and averral is related to the authors' purposes and reader orientation, but it is hoped that our discussion here has adequately illustrated the relationship.

## 8. Conclusion

In this paper I have attempted to show that text averral and attribution are important notions for our understanding of interaction in written text. I have also tried to show that the author's stated purposes and reader orientation are reflected in the pragmatics of text averral and attribution. One interesting observation is that attributions made in the introductory textbook are, with the exception of one, without citations. The reason suggested is that citations would weaken the authoritative voice of the textbook writer. In Bell, the author's purpose and reader orientation are different from those in Crane et al and this is reflected in the way attributions and text averral are manipulated. Since the author's stated purpose is to produce an integrated approach to the discipline, he is concerned to select what would serve his purpose best. This is manifest in the mitigating devices he employs in order to impress upon the reader that previous approaches are not to be discarded totally. In Sinclair and Coulthard the authors' purpose is to create a model of discourse analysis that would be adequate for the description of their data. This is evident in their review of the literature where naturally attributions and, more specifically, citations abound, the purpose being to show how previous models are all lacking and on what bases they are rejected.

Finally, from the pedagogical perspective, some useful observations can be made. It is very important to alert the students to the various voices they hear within the same text – to the signals of their onset and their offset. It is very important that they should be able to discriminate between the author's voice and the voices of those the author invites to take part in the creation of the text. There is a world of difference between:

> The statement of meaning is the weak point in language study.

and

> Bloomfield (1933) turned his back on the problem by observing that 'the statement of meaning is the weak point in language study, ...'

The first is averred, whereas the second is attributed.

The problem is that most students are unaware of the signals of text averral and attribution, and even in their own research papers they do not clearly signal when they have switched from expressing their own views to reporting or vice versa, with the result that they may be accused of at best ambiguity and at worst plagiarism.

## Notes

1. I am not claiming that these texts are typical of their disciplines – they just happen to fall within the convenient area of my professional involvement – nor are the pages that are analysed selected on any strict principles.

## References


Austin, J L (1962) *How to Do Things with Words,* Oxford: Oxford University Press

Bazerman, Charles (1981) 'What written knowledge does: three examples of academic discourse' *Philosophy of the Social Sciences,* 11. 361-82

Bell, Roger T (1976) *Sociolinguistics,* London: Batsford

Crane, L Ben; Yeager, Edward; and Whitman, Randal L (1981) *An Introduction to Linguistics,* Boston: Little, Brown and Co.

Cronin, Blaise (1981) 'The need for a theory of citing', *Journal of Documentation* 37. 1. 16-22

Francis, Gill (1986) *Anaphoric Nouns,* Discourse Analysis Monographs No. 11, Birmingham: English Language Research, University of Birmingham

Grice, H P (1981) 'Presupposition and conversational implicature' in Cole, P (ed) *Radical Pragmatics,* New York: Academic Press, pp 183-98

Harvey, Anamaria (1992) 'Science reports and indexicality', *English for Specific Purposes* 11. 2. 115-128

Hoey, Michael (1983) *On the Surface of Discourse,* London: George Allen and Unwin; reprinted in Reprints in Systemic Linguistics series, University of Nottingham.

Levinson, Stephen C (1983) *Pragmatics,* Cambridge: Cambridge University Press

Myers, Greg (1989) 'The pragmatics of politeness in scientific articles', *Applied Linguistics* 10. 1. 1-35

Myers, Greg (1992) 'Textbooks and the sociology of scientific knowledge', *English for Specific Purposes* 11.1.3-18

Sacks, H; Schegloff, E A; and Jefferson, G (1974) 'A simplest systematics for the organisation of turn-taking in conversation' *Language* 50. 696-735

Searle, J (1975) 'Indirect speech acts' in Cole, P and Morgan, J L (eds) *Syntax and Semantics 3: Speech Acts,* New York: Academic Press, pp 59-82

Sinclair, J McH (1983) 'Planes of discourse' in Rizvil, S N A (ed) *The Two-fold Voice: Essays in Honour of Ramesh Mohan,* Salzburg Studies in English Literature, Universität Salzburg

Sinclair, J McH (1985) 'On the integration of linguistic description' in Van Dijk, T (ed) *Handbook of Discourse Analysis,* Vol 2, London: Academic Press, pp 13-28

Sinclair, J McH (1986) 'Fictional worlds' in Coulthard, M (ed) *Talking about Text,* Discourse Analysis Monographs No. 13, Birmingham: English Language Research, University of Birmingham, pp 43-60

Sinclair, J McH (1988) 'Mirror for a text', *Journal of English and Foreign Languages* (Hyderabad, India) 1

Sinclair, J McH and Coulthard, R M (1975) *Towards an Analysis of Discourse,* Oxford: Oxford University Press

Swales, John (1981) *Aspects of Article Introductions,* Birmingham: Language Studies Unit, University of Aston

Swales, John (1990) *Genre Analysis,* Cambridge: Cambridge University Press

Swales, John (1991) 'The paradox of value: six treatments in search of the reader.' mimeo.

Tadros, Angele (1981) *Linguistic prediction in Economics text,* unpublished Ph.D thesis, University of Birmingham

Tadros, Angele (1985) *Prediction in Text,* Discourse Analysis Monographs No. 10, Birmingham: English Language Research, University of Birmingham

Winter, E O (1977) 'A clause relational approach to English texts: a study of some predictive lexical items in written discourse', *Instructional Science* 6. 1. 1-92

# THE CASE FOR THE EXCHANGE COMPLEX

*Michael Hoey*
University of Birmingham

## 1. Introduction

In 1975, John Sinclair and Malcolm Coulthard published what is now regarded as one of the key contributions to discourse analysis in the 70's: *Towards an Analysis of Discourse*, based upon an earlier extended SSRC report by John Sinclair et al (1972). In this book they argue persuasively for a structural approach to the description of classroom interaction and draw upon a rank scale model for the purpose which has as its basic unit the 'act' rather than any lexicogrammatical unit. Their approach has won both admirers and detractors, but its continuing usefulness is evidenced by the fact that papers modifying and expanding the model are still being published over twenty years after the original report; Coulthard (1992a) collects some of the more significant of these. This paper seeks to return to the original analogy with grammar and by slightly extending that analogy to make a further contribution to thinking about the structural status of interaction.

## 2. The rank scales of interaction and grammar

Sinclair and Coulthard (1975) draw heavily upon Halliday's early theoretical and descriptive work, the key concepts of which are to be found in Halliday (1961)[1]. According to Halliday (1961), grammar is organised in terms of a rank scale whereby words combine to make groups, groups combine to make clauses, and clauses combine to make sentences. The idea of the rank scale has proved robust and is reaffirmed in modified form in Halliday (1985), the main alteration being the replacement of the rank of 'sentence' with that of 'clause complex'; tagmemic linguistics also utilises the notion of the rank scale, though it does not use the term (Pike, 1967; Pike & Pike, 1977). Sinclair and Coulthard (ibid) argue that just as grammar can be described in terms of ranks, so also can classroom discourse. They posit the following scale:

Transaction
Exchange
Move
Act

Subsequently the possibility of another rank in between transaction and exchange has been posited, namely that of the sequence, noted in passing by Sinclair (1992).

This is defined by intonation and is quite different and separate from any proposal made in this paper (Brazil, 1985, Coulthard, 1992b).

To summarise Sinclair and Coulthard's model here in any detail would be unproductive, given that it is widely known and has been extensively described elsewhere. A thumbnail sketch, though, is obviously necessary. Sinclair and Coulthard see the exchange as the heart of classroom discourse. An exchange has three functional stages, an Initiation, a Response and a Feedback. Characteristically, a teacher will initiate, a pupil will respond, and the teacher will then provide feedback on the pupil's response; e.g.

| | | | |
|---|---|---|---|
| 1. | Teacher: | INITIATION | What do we do with a saw? Marvelette? |
| | Pupil: | RESPONSE | Cut wood. |
| | Teacher: | FEEDBACK | We cut wood. |

Each of these is realised respectively (and invariably) by an Opening, an Answering and a Follow Up move. For this reason, the labels for the functions and the type of moves have subsequently sometimes been used interchangeably. Coulthard and Brazil (1979) adopt the term Follow Up as a replacement label for Feedback which they find too limiting if Exchange Structure is to be used outside the classroom. In line with the original Sinclair and Coulthard model, the analysis of example 1 becomes:

| | | | |
|---|---|---|---|
| 1a. | Teacher: | INITIATION realised by Opening move | What do we do with a saw? Marvelette? |
| | Pupil: | RESPONSE realised by Answering move | Cut wood. |
| | Teacher: | FEEDBACK realised by Follow up move | We cut wood. |

The moves are realised by acts, one of which in each move will characterise that move and will be described as the head act; other acts either premodify or postmodify the head act and are optional. Just as a grammatical group may be realised by a single word, so a move may be realised by a single act. Thus both the Answering and the Follow Up moves in the above example are realised by a single act. The Opening move on the other hand is realised by two acts:

| | | | |
|---|---|---|---|
| 1b. | Teacher: | INITIATION realised by Opening move realised by an elicit act (a) and a nomination act (b) | What do we do with a saw? (a) Marvelette? (b) |

Transactions, the highest rank in the scale given above, are fairly shadowy units, distinguished more by their boundaries than by their internal construction. They are made up of boundary exchanges and teaching exchanges, teaching exchanges being exchanges of the kind just described. Like sentences, they are assumed to

participate in no higher structure. It will be the argument of this paper that the transaction is not the best unit to posit above the exchange. I shall return to this proposed rank near the end of this paper when I shall interpret it somewhat differently; in the meantime, it will not figure in my discussion.

The parallels between the rank proposed for classroom discourse and that used for grammar are quite close, and are best spelt out when the two ranks are placed next to each other:

| | |
|---|---|
| Clause | Exchange |
| Group | Move |
| Word | Act |

Some of the parallels are either inherent to the approach or adventitious. For example, one parallel is that words are unlimited in number and largely unanalysable, and that acts are likewise unanalysable and, if not unlimited, certainly indefinite in number (witness the various lists proposed by Sinclair and Coulthard's successors, e.g. Burton, 1980; Francis and Hunston, 1992). Another is that the structures of groups and moves are both amenable to analysis in terms of pre- and post-modification. A third is that grammatical description within the Hallidayan tradition has tended to give more attention to the structure of clauses than to that of groups (e.g. Sinclair, 1972; Halliday, 1985) and that likewise discourse description broadly within the Sinclair/Coulthard tradition has tended to focus on the structure of the exchange rather than that of the move (e.g. Coulthard and Brazil, 1979; Burton, 1980; Berry, 1981; Fawcett et al, 1988; Hoey, 1991). Sinclair (1992) makes exactly this point:

> The rank scale of act – move – exchange – (sequence) – transaction soon concentrated on the exchange, much as grammar was concentrating on the clause. (p.79)

Perhaps more significant than these points of similarity are parallels that suggest the force of the original analogy. To begin with, one element is essential to the functional analysis of a clause, namely the verbal group that realises the predicator or process of the clause (depending on whether our grammatical description is that associated with the interpersonal or ideational metafunction: see Halliday, 1985). Indeed long experience with students suggests that mistakes arise particularly often from misidentification of the predicator. In the same way, one element is essential to a functional analysis of the exchange, namely the initiation. If this is not identified correctly, little else can be correct.

A related parallel is that in transitivity analysis one lexical morpheme, that of the main lexical verb, determines the description of the whole clause; the only thing that distinguishes a nominal group functioning as actor from the same group functioning as sayer, senser and so on is the choice of lexical verb in the verbal group realising the accompanying process (Ravelli, 1991). Similarly, in exchange analysis, one act, the head act of the Opening move, affects the analysis of the remainder of the exchange; *O.K* may realise an acknowledgement, reply or react act in an Answering move depending upon the head of the Opening move.

Another parallel relates to the possible structures of clause and exchange. All indicative clauses have Subject and Predicator (Sinclair, 1972), or at least Subject and Finite with the Predicator ellipted (Halliday, 1985), but Complements (and

depending on one's terminology Objects) are not present in every clause. Likewise all exchanges have Initiation and Response but not all have Feedback (or Follow Up, again depending on whose terminology one uses)[2]. Indeed, as Burton discovered in her consideration of dramatic dialogue (1980), Feedback is uncommon in some interactive genres, while in others, like classroom discourse and quiz shows, it is virtually compulsory.

Lastly, some clauses are frozen, e.g. *all being well, if I were you, when all's said and done*. These clauses can be analysed as having clause structure but they function dynamically as single choices. The same is partly true, and perhaps more importantly so, for some exchanges. Exchanges such as the following are frozen:

2. A: Hi, there.
   B: Hi.
3. A: How are you?
   B: Fine. thanks.
4. A: Bye, then
   B: Bye.

Such exchanges are analysable as having exchange structure and indeed a separate choice is made by each of the contributors to the exchange in contradistinction to the lack of choices in the frozen clauses. Nevertheless the significance of these exchanges is that they are 'off the shelf' and that no processing is necessary in the act of producing them.

Although other parallels could have been drawn between the rank scales of the clause and the exchange, I hope enough has been done to demonstrate the usefulness of the original analogy of Sinclair and Coulthard's.

### 3. The absence of the exchange complex in the discourse rank

Although Sinclair and Coulthard make use of the rank scale in their description of discourse and show implicit awareness of points of comparison between discourse analysis and grammatical analysis (e.g. the use of pre- and post-modification at the move rank), there are several places in their description where they appear consciously to eschew close comparison of the levels of grammar and discourse. The first of these relates to the rank of sentence or, as Halliday now terms it, the clause complex. Just as words combine to make groups and groups combine to make clauses, so clauses combine to make sentences/clause complexes. Indeed Sinclair's own description of sentence structure (1972) is as full as any in the systemic tradition until Halliday's account of the clause complex in 1985. Yet there is no equivalent unit to the sentence/clause complex in Sinclair and Coulthard's model of discourse. The transaction is not an obvious candidate; as already noted, it is identified by its boundaries and no principles are suggested for the ordering or relating of teaching exchanges within it. It is also characteristically rather large, whereas a sentence may be realised by a single clause. There is therefore a gap where we might have expected, by analogy with the clause complex, an exchange complex to appear.

Clause complexes do not themselves participate directly in larger structures, but they are the units out of which texts are built. Because the exchange complex has not been posited, there has been no systematic exploration within the Sinclair-

Coulthard tradition of the possibility of such a unit forming 'interactive text', nor has it been suggested that exchanges might combine to form text. Ventola (1987), however, although not working within the Sinclair-Coulthard tradition, shows clearly that a structural description of interaction can be usefully combined with a description of the cohesive devices connecting the parts of the interaction; Tannen (1989) might be seen as indirectly contributing to such a description.

Any perceptive reader will by now have guessed (even without the help of the title) that a major purpose of this paper is to suggest that these missing parallels between discourse and grammar need not be missing. I wish to propose that the exchange rank be extended to include the exchange complex; I have reservations about the structural status of the transaction, which I shall air at the end of this paper. Part of the purpose of this paper is therefore to suggest the replacement of the transaction with the exchange complex in the rank scale. The arguments for sequence are unaffected by the proposal made here, though it is conceivable that it might be reinterpreted as a 'compound exchange' (See section 5.4). The rank I shall argue for is therefore as follows:

Exchange complex
Exchange
Move
Act

I want further to argue that exchange complexes function to create interactive texture by means of a special form of thematic development separate from but operating on related principles to those utilised by Halliday to account for the theme-rheme relations he describes in connection with the textual metafunction (1985).

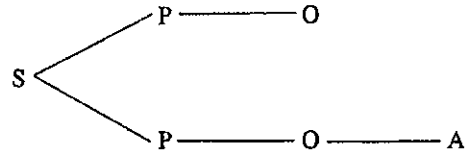
## 4. Some features of the sentence/clause complex

So that we know what features we might look for in the proposed exchange complex, it is worth briefly reminding ourselves of the ways in which simple clauses may combine to form a more complex structure. All the examples in this section are unauthentic since I seek to represent a grammatical consensus rather than argue for or against the appropriateness of particular structural descriptions; when we come to look at exchange complexes, on the other hand, all examples will of course be genuine.

In the first place, clauses may combine in branched structures; e.g.:

5. I bought a paper and took it home.

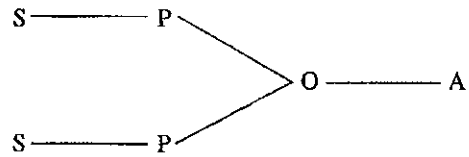
This can be represented thus:




They may also (more rarely) combine in converging structures; e.g.:

6. His mother earnt and his father squandered a small fortune each month.

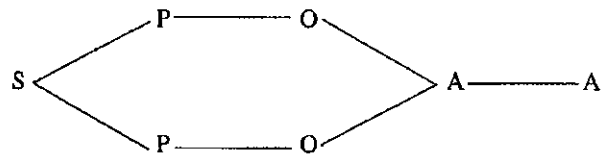
This can be represented in a similar fashion:




Logically these two possibilities combine to make a third possibility: the branched and converging structure; e.g.:

7. She prunes the roses and mows the lawn without fail every Sunday;

which can be represented as follows:




Perhaps, though, the most characteristic way in which clauses combine is by means of subordination. For the purposes of my argument here the variety of kinds and meanings of subordination is not relevant. All that needs to be noticed is that we can have a Free clause followed by a Bound clause which, as Sinclair's term (1972) implies, is structurally tied to the free clause; e.g.:

8. I bought a paper (F) so that I could choose which film to go to (B).

We can also have a Bound clause within a Free clause, notationally represented by Sinclair as F[B]; e.g.:

9. The argument ($F^1$), although it became heated ([B]), was never abusive ($F^2$).

Bound clauses are always structurally dependent on Free clauses.

Finally, clauses may combine to make compound sentences. This is the loosest kind of structural expansion and occurs when two Free clauses are linked by a coordinator; e.g.:

10. First I bought a paper and then I went to the cinema.

It is not always easy to tell in such circumstances whether one has a single clause complex or two simple clauses.

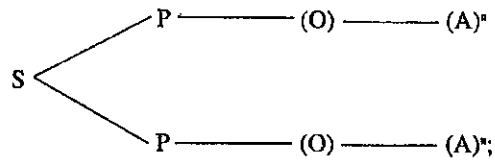
## 5. The exchange complex

In this section we will explore the extent to which the analogy between the posited exchange complex and the sentence/clause complex may be said to hold. We should not expect the parallels to be exact, since we are comparing possibilities on different levels, using different units which occur under quite different circumstances of production; the sentence/clause complex is produced in comparatively little time by a single speaker/writer who has complete control over the whole, whereas the posited exchange complex will be produced by at least two speakers, neither of whom will have the same control over the whole, and it will take a noticeable amount of time to produce. Indeed so obvious is it that the initiator of

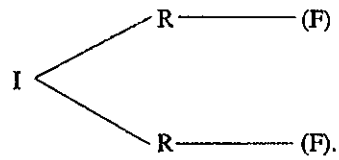
an exchange cannot know how it will end up that a number of writers have attempted to represent interaction dynamically (e.g. Fawcett et al, 1988, Ventola, 1987, Hoey, 1991). But that is no reason for not pursuing the comparison. Since clause complexes have been shown to be representable dynamically (Nesbitt and Plum, 1988) and even the structure of clauses have been shown to be amenable to dynamic description (Ravelli, 1991), it seems reasonable to reverse the process and examine the data of interaction synoptically. Ravelli indeed argues that all data are describable in either dynamic or synoptic terms and that it is not the case that some kinds of data are more suited to the one kind of description than the other.

*5.1 Branched structures*

Given that the clause complex may have variants derived from the pattern




we are looking for interactions that have the pattern

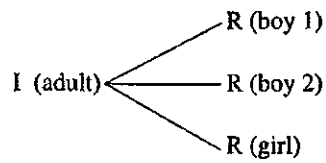



The following interaction seems to illustrate one of the possibilities:

11. Adult: What about if we introduced the word *house husband*? Would that be a good idea?[3]
    Boy 1: No way!
    Boy 2: No
    Girl: Yes 'cos some of the ladies might like to go out to work instead of staying home and looking after the children.

(data collected by Dawn Morris)

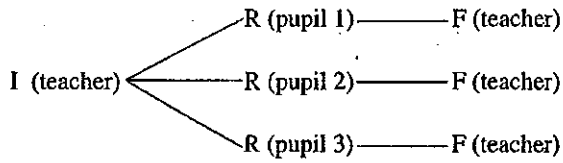
To the two boys the exchange is a simple one; there is no evidence that the second boy is interested in reacting to the first boy's Response. The girl on the other hand may be aware that she is contributing to an exchange complex in that her utterance is interpretable as being a riposte to the boys at the same time as it is a Response to the adult's Initiation. Nevertheless, I would suggest that the following analysis of 11 is possible:




Now consider the following interaction from Sinclair and Coulthard's own data:

12. Teacher: What's the name of each of those?
Pupil 1: Paper clip
Teacher: Paper clip
Pupil 2: Nail
Teacher: Nail
Pupil 3: Nut and bolt
Teacher: Nut and bolt.

Each pupil is responding separately to the original (multiple) Initiation. A reasonable analysis would seem to be:




Although we are not privy to the intonation used in the final teacher utterance it seems likely that it marked the end of a series. If so, then the intonation would be serving a similar function to the final coordinating *and* in a series of clauses, bringing the exchange complex to a recognisable end.

Sinclair and Coulthard (1975) discuss example 12 but in different terms. They treat it as a kind of Bound exchange (of which more below). They note:

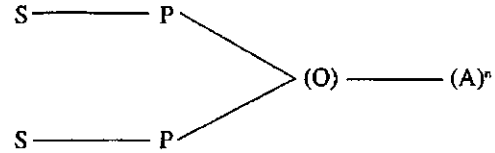
> Occasionally teachers withhold evaluation until they get two or three answers. Sometimes they are making sure that more than one person knows the answer, sometimes they have asked a multiple question. In this case the structure is... IRF(Ib)RF(Ib)RF...Ib is only realised by nomination and the F preceding Ib contains no evaluation. (p.55)

This analysis seems less satisfactory than our own. Firstly, as we shall discuss in more detail in section 5.3, the status of a Bound exchange is here being left unclear. It is difficult to tell whether we should treat it as a separate exchange and therefore a unit of the same kind and status as other exchanges around it, or whether it should be regarded as being structurally tied to the previous exchange in which case we need a rank unit above the exchange and below the transaction. Secondly, we are required to assume elipted bound Initiations and yet we are not invited to draw the obvious conclusion: that an exchange dependent upon a previous exchange is structurally bound to it in exactly the same way that a clause with an ellipted subject is structurally bound to its predecessor.

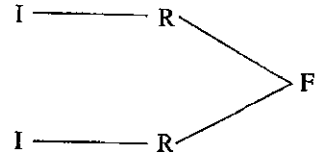
The problems disappear if we assume a rank above the exchange. Examples 11 and 12 suggest that branching occurs in interaction as in grammar and that when it occurs we have one kind of exchange complex.

### *5.2 Converging structures*

We saw that the clause complex may have variants derived from the pattern



where ( ) indicate alternatives, one at least of which must be chosen. We are therefore looking for examples that could be derived from the pattern:



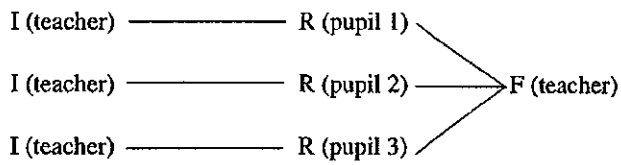

Because of Sinclair and Coulthard's restriction of exchange structure functions to three, no alternatives are possible. Obviously if Coulthard and Brazil's (1979) or Stubbs' (1983) suggestions for additional exchange functions were to be adopted, the pattern of converging could be more complicated.

An example that appears to fit the above pattern is the following; a mathematics class is in progress and graphs are being examined:

13. Teacher: Can anyone tell me what is halfway between 400 and 500?
    Pupil 1: 450
    Teacher: What is the quarter way mark between 400 and 500?
    Pupil 2: 425
    Teacher: What is the three quarter way mark?
    Pupil 3: 475
    Teacher: So it would be 425, 450, 475, 500. Do you get the idea?

(data collected by Susan Rees)

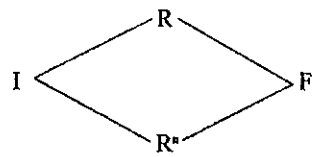
The Feedback unambiguously provides evaluation of all the previous Responses in the extract, not just the final one. So it seems reasonable to analyse it thus:




If this is accepted, a second kind of exchange complex may occur as a result of converging.

*5.2.1 Co-occurring branched and converging structures*

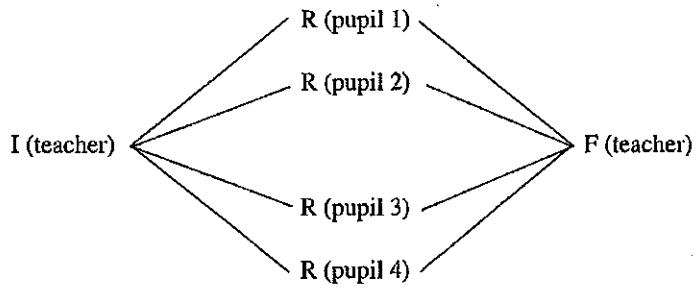
As was noted in section 4, branching and convergence can co-occur in grammar. There is only one pattern of co-occurrence that is possible in discourse, given the restriction to three exchange functions, and that is:




The following extract from the same data as the previous example seems to illustrate such structuring:

14. Teacher: And even more so, in the year 2000 we don't know, but we can...?
    Pupil 1: Guess
    Pupil 2: Imagine
    Pupil 3: Predict
    Pupil 4: Estimate
    Teacher: Yes, imagine, guess, predict, estimate, they all mean the same

Again, the Feedback unambiguously evaluates the answers as a group, rather than individually. Thus example 14 would seem to conform to the following structure:




Obviously other analyses are possible – we could argue that a teacher pointing at a child counts as a separate Initiation, though we would have to accept that the act of nomination can function as the head of the move that realises such a non-verbal Initiation – but it is as simple to assume that exchange complexes may be attested in the form of co-occurring branched and converging structures.

*5.3 Subordinate exchanges*

The most common form of clause complex involves subordination. So we should expect to find that a similar phenomenon to grammatical subordination is operating to create exchange complexes if the analogy with the clause complex is valid. In this section we argue that in fact the notion of subordination has been lurking in the model from its inception. We shall show that it has always been recognised that exchanges structurally combine in such a way that the later exchanges make no sense unless they are seen as attached to the first exchange; our case is that this is a kind of subordination. Oddly, Sinclair and Coulthard appear to have anticipated my discussion here without drawing the obvious conclusion. They note five kinds of Bound exchange (one of which we have already discussed in section 5.1), commenting:

> An exchange is bound either if it has no initiating move, or if the initiating move it does have has no head, but simply consists of nomination, prompt, or clue. (p.53)

The decision to label such exchanges Bound would seem to invite comparison with the notion of the Bound clause (Sinclair, 1972), and yet the question is not raised of whether the Bound exchange and the Free exchange form a larger unit. Consequently, as noted in 5.1, we are left in doubt as to whether we should treat the Free exchange/Bound exchange as a variant form of exchange or as two separate exchanges. In Hoey (1991) I attempted to describe a variety of ways in which

an exchange might be extended. At least two of these might be considered to be kinds of subordination. The first of these was noticed and discussed by Stubbs (1983). He cites the following example:

15. A: Can you tell me where the Savoy Cinema is?
    B: Ooh yeah it's only round the corner here
    A: Is it?
    B: It's not far like
    A: Cheer thanks very much ta

Of this example he notes that A's second utterance is problematic because it is a non-initial Initiation. His solution is to posit a further exchange function *Ir* (Re-initiation), so that exchanges may have the pattern I R Ir R with Ir R being recursive. (The term Re-Initiation is taken from Sinclair and Coulthard's discussion of Bound Exchanges, although Stubbs uses the term more broadly.) Given the heavy ellipsis in A's second utterance in example 15 it can be argued that this is in fact an exchange complex with the second exchange subordinate to the first. The important thing to notice about Stubbs' example is that there is no new information contained in A's second utterance at all. It does not offer a new topic or significantly modify the current one. We will only consider an utterance to be a Re-Initiation if there is no new topic offered; mere unintelligibility out of context will not suffice.

Some further examples may help indicate the way Re-initiation may operate. In the first of these examples, a quizmaster is showing contestants photographs of people who were famous in a particular year:

16. Quizmaster: Number two?
    Contestant: I think that was the space woman, Valentina somebody or other er umm
    Quizmaster: Give me a stab at something or { other
    Contestant: { Tetra.. Terra
    or..
    Quizmaster: Say something in Russia
    Contestant: Terraskoffa or { something
    Quizmaster: { Yeah you got it

(data collected by Karen Brown)

The brackets indicate momentary simultaneous speech.

In this example we have two instances of Re-Initiations (at least as we are using the term), namely the quizmaster's second and third utterances (**Give me a stab at something or other** and **Say something in Russia**). They gain all their point from being attached to the first exchange which they seek to bring to a happy conclusion. Thus our analysis would be

FREE I (quizmaster) R (contestant)
|
BOUND Ir (quizmaster) R (contestant)
|
BOUND Ir (quizmaster) R (contestant) F (quizmaster).

One difference between grammatical subordination and exchange subordination is that the final Feedback/Follow Up brings the whole complex to a conclusion. Components of Bound clauses do not directly affect the grammar of the Free

clause to which the Bound clauses are attached. Notice that a Re-initiation cannot occur after the Feedback function.

Another example is the following piece of classroom data, collected by Rachel Lopata:

17. Teacher: What was the name of the boy who found Oliver and took him back to Fagin? Come on.
    Pupil: Dodger
    Teacher: He was called?
    Pupil: Dodger
    Teacher: Dodger, yes. And Oliver ended up working for Fagin as a pickpocket. And that was a story of a little boy who lived during Victorian times.

The teacher's second utterance, asking for repetition presumably so that the whole class can hear, introduces no new topic nor modifies an existing one. It is in our terms therefore a Re-Initiation and the exchange which it begins is a Bound Exchange:

FREE I (teacher) R (pupil)
|
BOUND Ir (teacher) R (pupil) F (teacher)

Example 16 seems to have much in common with Sinclair and Coulthard's first class of Bound exchange (Re-initiation 1), though it is not identical with it. Sinclair and Coulthard comment:

> When the teacher gets no response to an elicitation he can start again using the same or a rephrased question, or he can use one or more of the acts – prompt, nomination, clue – to re-initiate. The original elicitation stands and these items are used as a second attempt to get a reply. (Sinclair and Coulthard, 1975, p.53)

The cases Sinclair and Coulthard are attempting to cover are those where there is no Response. But in other respects they describe the situation illustrated by example 16. In section 5.5 we will see a case where Ir is by new nomination.

Example 17 is an example of their fifth class of Bound exchange (Repeat). While it may be important on occasion to separate different types of Ir, the important point is that Ir can be treated as a marker of subordination just as much as a subordinator or *which/who* etc. Of course the meaning of the subordination is different and the structural possibilities are consequently also different. The possibility available for contingent bound clauses to be mobile in the sentence/clause complex is not available for the Bound exchange. But then neither Reported Bound clauses nor Adding Bound clauses have such structural freedom either and we do not deny these clauses subordinate status on such grounds.

A second class of what we might consider a subordinate exchange is triggered by a feature of the Free exchange, namely the nature of the Feedback. Whenever the Feedback is negative, i.e. whenever the fit between Initiation and Response is evaluated as in some ways inadequate, incorrect, inaccurate or inappropriate, the negative Feedback functions as a Re-Initiation. This is a comparable situation to that described for written discourse in Hoey (1983) where a text may have the pattern Situation – Problem – Response – Evaluation, but where a Negative

Evaluation triggers off a further Response and Evaluation. In both interaction and written discourse the pattern continues to recycle until a positive Feedback is achieved. The interactions that follow are examples of negative Feedback triggering off further Responses.

18. Husband: What's the time?
Wife: I don't know. About ten o'clock I think
Husband: We don't want to miss the Woody Allen
Wife: No *[she leaves the room to look at clock; returns]* It's ten past ten

(own data)

19. Quizmaster: What is the name of Barry Humphries' Minister for cultural affairs? Nigel?
Nigel: Les
Quizmaster: Les what?
Gary: Dawson
Quizmaster: No *[laughs]*
Tony: Patterson
Quizmaster: Patterson...Tony Blackburn jumped in.

(data collected by Karen Brown)

In both these cases, the Feedback can be interpreted as evaluating as inadequate or incorrect the fit between the Initiation and the Response. In the first case, the husband's second utterance evaluates the wife's rough estimate of the time as inadequate for his purpose, namely to ensure that they catch the beginning of a Woody Allen film on TV. In the second case the quizmaster's third utterance and laughter evaluate the previous Response as incorrect. Sinclair and Coulthard are aware of such cases, which they handle as their second class of Bound exchange, though they do not specifically relate the boundness to negative Feedback; again, of course, the issues as to the exact status of Bound exchanges arise. My argument would be that we have in these interactions instances of the exchange complex, and my analysis of 18 would be:

| | | | | | |
|---|---|---|---|---|---|
| FREE | I | (Husband) | R | (wife) | NegF (Husband) |
| \| | | | | | |
| BOUND | | | R | (wife) | |

and of 19:

| | | | | | |
|---|---|---|---|---|---|
| FREE | I | (quizmaster) | R | (Nigel) | |
| \| | | | | | |
| FREE | I | (quizmaster) | R | (Gary) | NegF (quizmaster) |
| \| | | | | | |
| BOUND | | | R | (Tony) | PosF (quizmaster) |

The second Initiation by the quizmaster is not regarded as Bound in the sense in which we are using it, despite its heavy dependence for intelligibility on the previous exchange (and despite Burton's (1980) use of the term in this way). The reason is that it asks a new, though obviously closely related, question, and its relationship to the previous exchange is the same as that of a new clause to its predecessor when there is heavy cohesion; e.g.:

20. They thought he would fail. He did.

It will be argued in section 6 that exchange complexes contribute to the creation of 'interactive texture'. Although cohesion will not be specifically discussed, it should be apparent from that discussion that exchange complexes – and particularly the initiations of those complexes – are often linked by cohesive devices of various kinds, particularly ellipsis.

Although the quizmaster's second Initiation does not mark subordination, the negative Feedback does, and this is therefore another way in which an exchange complex may be formed. Occasionally, though, the boundaries between this and the previous kind of subordination blur. If for example in a classroom a pupil gives an incorrect Response and the teacher without commenting nominates a new speaker, it will probably be heard by the first pupil as implying negative Feedback, although in our terms it would be an instance of Re-initiation.

We have been considering two kinds of connection between exchanges that seem to justify use of the term 'subordination'. Other kinds of 'subordination' could be posited. For example, in our brief discussion of clause subordination we saw that a Bound clause may be 'embedded' within a Free one: F[B]. I would suggest that the phenomenon handled by Jefferson (1972) as 'side sequences' could be handled as a form of embedded Bound exchange, making use of the notion of Challenging, as described by Burton (1980) and adopted by Sinclair (1992), or of Counter-Initiation as described by Longacre (1976). Space, however, does not permit the working out of this possibility in detail.

To summarise our argument throughout the whole of this sub-section, if we take seriously Sinclair and Coulthard's notion of the Bound Exchange and take their choice of label for these exchanges at face value, subordination can be said to exist within the model and should be accepted for what it is, a structural relationship binding one exchange to another in such a way that one of them is felt to be interpretable only in terms of the other.

### *5.4 Compound exchanges*

In section 4, we noted that the final way in which clauses might combine into sentences/clauses complexes was by means of coordination without ellipsis, resulting in what is often referred to as the 'compound sentence'. There are two ways in which an equivalent structural relationship of coordination between exchanges might be identified. The first is by means of intonation, in which case the posited rank of 'sequence' would be the same as the posited compound exchange. Since I have not researched this possibility, I do not discuss it further here.

A second way in which coordination might be identified between exchanges is by separating out the different participants' varying perceptions of, and consequent behaviour in, the interaction. Most of the time in classroom discourse the analyst behaves as if there are no discrepancies between the teacher's and the pupils' perception of the interaction. But in other kinds of interaction, and at times even in classroom interaction, there may be grounds for recognising slight but significant potential differences between participants. Specifically, what may be offered by one participant as a Feedback may be interpreted by the other(s) as an Initiation or what is offered as a Response may be heard as simultaneously functioning as an Initiation. Consider the following example:

21. A: Did you see that programme last night on the TV?
    B: No, I haven't got a telly.
    (data collected for Eija Ventola as part of course taught by EV and MH)

B's utterance seems uncomplicatedly to be functioning as a Response to A's Initiation and it would have been open to A to provide a Feedback such as 'pity.' The way A continues is however as follows:

21a. A: Don't you? Haven't you got a telly?

This utterance is treating B's Response as if it were an Initiation and is performing a Challenge function (Burton, 1980, Sinclair 1992). (The speaker is expressing doubt about whether B is conforming to Grice's (1975) Maxim of Quality). Since what we have are two exchanges with a shared utterance, one possibility is to treat this as a kind of compounding:

Initiation (A) ——— { Response (B)
                    { Initiation ——— Challenge (A)

The second Initiation has not been attributed to B even though the utterance is incontrovertibly B's since we have no grounds for believing that B intends to initiate. The point is that A chooses to treat it as if it were an Initiation; for all we know, A may not believe B intended an Initiation.

Notice that what separates such a case from the instances of subordination we were earlier considering is the fact that both these exchanges are interpretable on their own terms; neither of the exchanges is dependent on the other for its interpretation, at least no more so than is normal between exchanges. It is this relative independence of the combined exchanges that permits us to treat such a combination as an instance of coordination rather than subordination.

Francis and Hunston (1992) quote the following example; the conversation centres on a Helium balloon which has mysteriously stopped floating:

22. B: Why i-it's not floating at all?
    A: No it's lying on the floor like any old balloon
    B: It's a bit *strange* you know
    A: Yeah interesting

Francis and Hunston's analysis of this is that there are two exchanges: IR, IR. But this analysis assumes that both parties knew what each other were going to say. If we didn't have A's final utterance we would have had no grounds for not interpreting B's second utterance as providing the expected evaluative Feedback after A's Response. I suggest therefore that this is a case where what was offered as Feedback by one of the participants has been treated as if it were an Initiation by the other. In accordance with our discussion above, therefore, our analysis would be:

Initiation (B) —— Response (A) —— { Feedback (B)
                                   { Initiation —— Response (A)

Again, the two exchanges are reasonably self-standing; again, then, the analogy with coordination rather than subordination seems more apt. The analysis just given is not in fact a novelty, Wells et al (1981) talk of such exchanges as overlapping, and Francis and Hunston consider analysing example 22 in exactly this way. Their reason for rejecting such an analysis is significant:

> Analyses of this type satisfy the intuition that exchanges are linked together in a way that rigidly-defining horizontal lines between them tend to mask. Such a solution, however, wreaks havoc with a hierarchical system of analysis. For the exchange to be a unit which will combine with other exchanges to form larger units, or transactions, it must have clearly defined boundaries and there must be limits on what it can contain. Otherwise it is impossible to apply the notion of rank scale to ordinary spoken discourse. (Francis and Hunston, 1992, p 151)

Francis and Hunston's problem is that the exchange is being made to carry too much weight. If the exchange is both the simple IRF structure and a complicated web of overlapping IRF patterns, then of course the hierarchical system of analysis is threatened. But the problem disappears if we posit the unit of the exchange complex. Then the larger unit of the transaction – if it exists – will be made up of exchange complexes, not directly of exchanges. An exchange complex, we have seen, may be created by branching, converging, subordination, coordination or any combination of these, though, as at the other ranks, it is perfectly possible for an exchange complex to consist of a single exchange. One would expect the variety and types of complexity to be affected by the situation. Thus classroom discourse may, for example, favour uncomplicated exchange complexes, while casual conversation may favour highly elaborate ones.

### *5.5 The exchange complex: a final example*

For obvious reasons the examples given so far have shown just one kind of exchange complex structure at a time. But just as sentences may show several kinds of complexity at once, so also may the exchange complex. In this subsection we conclude our discussion of the internal structure of the exchange complex by discussing a slightly more complicated example taken from the same fairly relaxed quiz programme as examples 16 and 19. Lettering has been added for convenience of reference:

23. Quizmaster: ..(a) and as a prisoner he too had a number, what was it?
Moira: (b) Well he didn't have a number because he was for freedom.
Quizmaster: (c) Ahh fear not, no
Moira: (d) Four?
Quizmaster: (e) Nigel?
Nigel: (f) We kind of like ⎧ six
Gareth: (g) ⎩ six
Quizmaster: (h) Six ⎧ it was
Moira: (i) ⎩ ahhhh
Quizmaster: (j) Yes it was

A number of devices for exchange combination are in operation here:

(i) The quizmaster's Feedback (c) to Moira's Response is negative, setting up a likelihood that she will feel obliged to offer another Response (given that the rules allow it here) which she duly does. This then combines the first and second exchange (comprising a-d): [IRNeg F]-[IR], where the dash indicates subordination and the connecting line represents a shared element and double coding. Each bracket represents a component exchange.

(ii) Instead of providing Feedback to Moira's Response the Quizmaster provides a new Initiation by means of nomination. This was one of the ways in which Re-initiation could occur and has the effect of combining the second and third exchanges (comprising c-h): [IR]-[IrRPosF].

(iii) Gareth and Nigel answer simultaneously (f and g). Since both their utterances are appropriate as Responses, they can be seen as an instance of branching, albeit one governed by simultaneity rather than sequence, and are followed by converging. This can be represented thus: [IR]b[R]c[F], where b indicates the branching of the adjacent functions and c indicates the convergence of previous elements with the following element.

(iv) Moira's reaction (i) to the Quizmaster's Feedback (h) can be seen as an example of compounding. She treats his Feedback as if it were an Initiating inform and provides a Response in the form of an acknowledgement; the latter is sufficiently vocal to provoke the quizmaster into providing Feedback. Thus the final exchange (h-j) is coordinated with the ones before as follows: [IR]b[R]c[F] [IRF]. As before the connecting line indicates a shared element and double coding; since we have suggested no way in which a compound exchange complex could occur that did not involve a shared element, there is no need to separately mark the compounding.

The total analysis of the single exchange complex that makes up example 23 is therefore as follows:

[IRNegF]-[IR]-[IrR]b[R]c[PosF] [IRF]

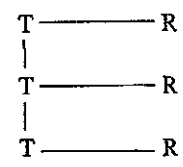
Obviously there are questions as to how many exchanges a branching and converging structure is made of, but even if this aspect of the analysis is discounted we have here an exchange complex that is four exchanges long. Presumably further research will uncover still larger and more intricate exchange complexes.

## 6. The exchange complex's role in creating interactive texture

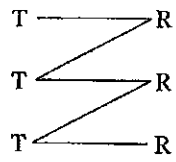
Clause complexes/sentences are the highest unit in the grammatical rank, at least in the systemic tradition, and their relationship with each other is not structural but textural, as argued by Halliday and Hasan (1976). One of the ways in which they relate to each other is through theme-rheme development (Daneš, 1974). Exchange complexes, it is argued in this section, also relate to each other thematically, though the development is of a slightly different kind. To understand what exchange complexes share with clause complexes/sentences and where they differ, it is necessary to outline briefly the kinds of theme-rheme pattern that may occur across sentences/clause complexes.

Daneš notes the following four patterns of theme-rheme development:

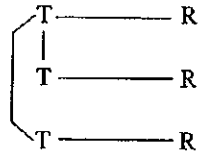
(1)



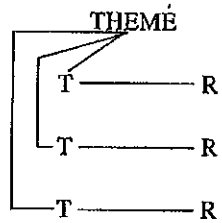
(2)




(3)




(4)

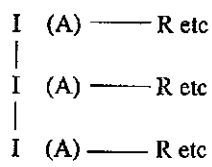



The first represents the situation where each new sentence/clause complex takes its theme from the theme of the previous sentence/clause complex. The second represents the situation where each theme is drawn from the previous rheme. The third reflects the situation where each theme is drawn from an initial theme. The last recognises a situation where there is a macrotheme from which all the themes are derived.

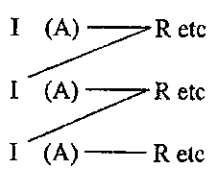
These four patterns do not exhaust all the possibilities, as Lau (1992) has shown, but they give a reasonable idea of the range of possibilities. In ordinary writing the various patterns of theme-rheme development merge so that an analysis of any particular piece of prose may present a much messier picture than any of these idealised patterns suggest. Nevertheless as a way of talking about theme-rheme development they are very valuable.

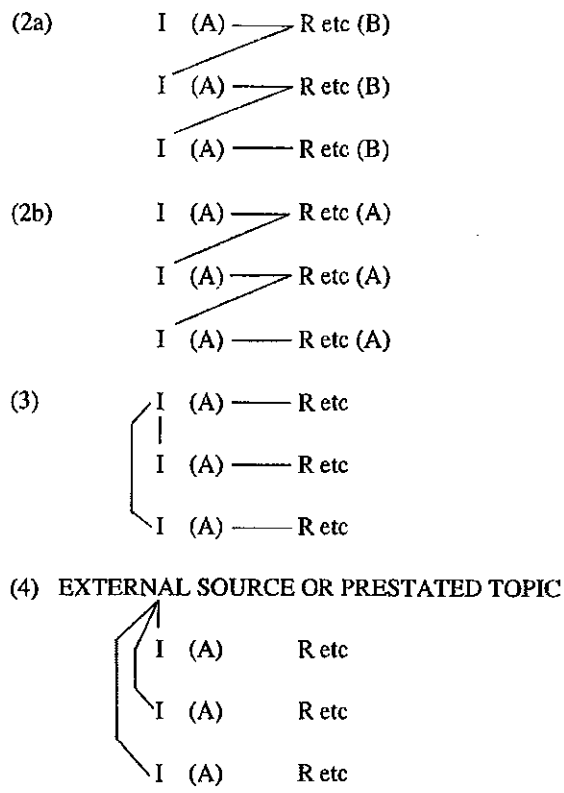
A parallel situation exists for the exchange complex. If we treat the initial Initiation as Theme-like and all subsequent elements in the exchange complex as Rheme-like, it is possible to posit that exchange complexes will manifest interactive patterns like (but not of course the same as) those listed above. Patterns that might be found are the following; A and B represent the participants:

(1)




(2)






The first of these possibilities describes a situation in which each exchange complex initial Initiation draws its content from the previous one. The second pattern describes a situation where each exchange complex initial Initiation draws on some utterance or utterances in the 'rheme' of the exchange complex. Patterns 2a and 2b describe the same situation except that the new Initiation draws on something that *either* B *or* A said in the 'rheme' of the previous exchange complex. Characteristically, in classroom discourse, 2a would occur when the teacher initiated on the basis of what one or more pupils had said in the previous exchange complex, while 2b would be what would happen if a teacher's exchange complex initial initiation was drawn from his or her own Feedback in the previous exchange complex. Pattern 3 describes a situation where several exchange complex initial Initiations draw on the content of a key earlier one. The final pattern covers cases where an overall topic is stated explicitly at the outset or where the Initiations derive from an external source (e.g. a map, a questionnaire). These patterns should be seen as representative not exhaustive.

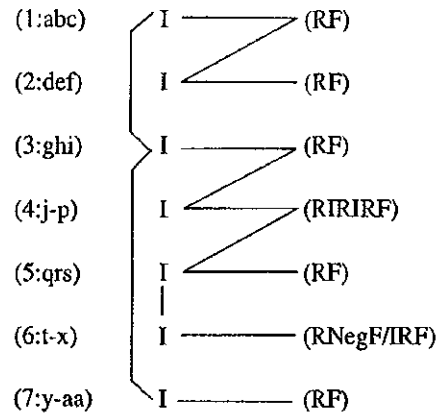
Just as most texts display a mixture of T-R patterns in their thematic development, so most interactions display a mixture of the above I-R patterns in their interactive development. But the following example should show how the notion works for simpler cases. The interaction is analysed first in terms of the modified form of Sinclair/Coulthard model proposed here and then the interactive development is mapped out. The example to be analysed is part of a class in which there is team teaching between a subject specialist and a language specialist. Except where named, pupils are not distinguished; it should not be assumed that there was just one very bright pupil who answered all the questions! Lettering has again been added for convenience of reference:

24. Subject teacher: (a) What does the 'y' axis tell you? It tells you something else.
Pupil: (b) Human population
Subject teacher: (c) Human population. (d) And what is human population measured in?
Pupil: (e) Millions
Subject teacher: (f) Millions, right. (g) Now what I'd like you to do is look at the numbers. What does the 'x' axis go up in every time?
Pupil: (h) In 100's
Subject teacher: (i) Good. (j) Can anyone tell me what is halfway between 400 and 500?
Pupil: (k) 450
Subject teacher: (l) What is the quarter way mark between 400 and 500?
Pupil: (m) 425
Subject teacher: (n) What is the three quarter way mark?
Pupil: (o) 475
Subject teacher: (p) So it would be 425, 450, 475, 500. Do you get the idea?
Language teacher: (q) Does Kalsuma understand that? What's half of a hundred, Kalsuma?
Kalsuma: (r) 50
Language teacher: (s) Yes, 50, good. (t) What's three quarters of 100, Kalsuma?
Kalsuma: (u) 25
Language teacher: (v) *Three* quarters, Kalsuma
Kalsuma: (w) 75
Language teacher: (x) 75, good. Good
Subject teacher: (y) What does the 'y' axis go up in every time?
Pupil: (z) 200's
Subject teacher: (aa) In 200's, yes, Good.

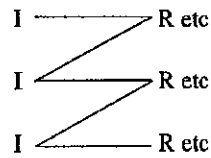
(data collected by Susan Rees)

We can recognise six exchange complexes here. The first is made of just one exchange and comprises a, b and c (the exchange structure being IRF); the second is likewise made of one exchange (d, e and f: IRF), as is the third (g, h and i: IRF). Having an exchange complex consisting of only one exchange is no odder than having a sentence made of only one clause. The fourth exchange complex, which contains three exchanges, was discussed earlier and comprises j – p; its structure is [IR][IR][IR]c[F]. The fifth contains only one (q, r and s: IRF), but the sixth contains two exchanges connected by subordination, v being an example of negative Feedback; it comprises t – x and its structure is [IRNegF]-[IRF]. The final exchange complex is again made of just one exchange (y, z and aa: IRF).

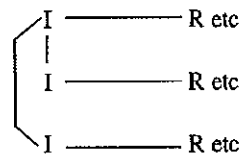
The organisation of the interactive thematic development of the interaction is as follows:




Since one or other teacher is the exchange complex initiator in every case, I have left participants out of the above display. It will be seen that two patterns predominate. The first is pattern 2:

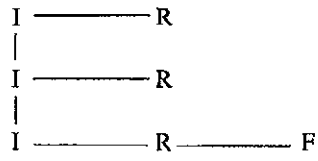



and seems to correlate with the pursuit of closely related topics. The second is pattern 3:

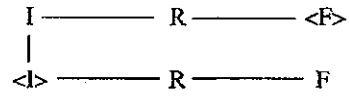



where each new exchange complex initiation takes its cue from an important earlier one. This occurs twice in our analysis of the classroom interaction. In exchange complex 3, Initiation g, with its topic of the 'x' axis, looks back to the first Initiation (a) with its topic of the 'y' axis. Likewise, in exchange complex 7, the subject teacher's Initiation (y), which looks at the scale of the 'y' axis, in part draws its content from her earlier exchange complex Initiation (g) back in exchange complex 3, which looked at the scale of the 'x' axis. This pattern of interactive development seems to occur whenever the teacher wishes to 'move the subject on'.

Just as theme-rheme analysis can be applied either to sentences/clause complexes as a whole or to the individual clauses, so also interactive development may be applied either to the exchange complexes as a whole, as I have just done, or to the individual exchanges. Applied to the internal exchanges of exchange complexes 4 and 6, we find in both cases the same pattern, namely pattern 1:




and




(where the angular bracketing indicates identity). This suggests that, at least for the classroom in question, interactive development at the level of the exchange complex is according to patterns 2 and 3 and within the exchange complex is according to pattern 1.

What then of the transaction? The implication of the fact that the exchange complex is involved in the creation of interactive texture is that the exchange complex is the highest *structural* unit of interaction. But this leaves us with the sequence and the transaction to account for. That the sequence (Brazil, 1985) is a structural unit I do not question, but its place is probably in a different rank as Brazil notes. The transaction is better regarded not as a structural unit (i.e. with internal rule-governed organisation) but as an organising unit like the paragraph. In Hoey (1985) it is argued that paragraphing reflects the judgements of the writer about the way chunks of text relate together. In other words, the paragraph is defined not in terms of its internal characteristics but in terms of the relations it forms with other paragraphs. A paragraph indentation is a sign of boundary and of relation and/or difference between chunks of text. In the same way, I would argue, the transaction is not defined in terms of its internal structure – which, apart from the so-called boundary exchange(s), is simply an unordered sequence of exchanges (or in our terms exchange complexes) – but in terms of the relations it forms with other chunks of interaction. In other words, when a controlling participant in an interaction indicates a transaction boundary, s/he is indicating that the current 'chunk' of interaction relates as a block to other 'chunks'; s/he is not indicating that the subsequent interaction will conform to any structural principles. So we accept the reality of transactions as organising features of classroom discourses and occasionally other kinds of discourse, but reject its structural status. Our rank for classroom (and other) discourse is therefore:

exchange complex
exchange
move
act.

## 7. Conclusion

This paper has argued that discourse structure as first laid down in Sinclair and Coulthard (1975) should be modified to allow an extra unit above the exchange – the exchange complex. By analogy with the clause complex or sentence, it has been argued that exchange complexes may be made up of branching exchanges,

converging exchanges, subordinate exchanges and coordinated exchanges, or any combination of these. Positing its existence allows one to remove anomalies in the current description and to explore the interactive development of a discourse. It is a measure, however, of the value of Sinclair and Coulthard (1975) that eighteen years on it is still worth while crossing their 't's.

## Notes

1. Sinclair and Coulthard have been criticised for making less use of the *systemic* aspects of Halliday's work characteristic of his later thinking e.g. Fawcett et al (1988).
2. I do not accept that an Initiation with an inform act as the head of the Opening move is not obligatorily followed by a Response (Hoey, 1991). In my view, the mistake here is to assume that the slots of an exchange must be sequential only, rather than either sequential or simultaneous. If we allow that Initiation and Response can be realised simultaneously, then the phenomenon described as backchannelling (e.g. Duncan and Fiske, 1977) can be seen as realising a simultaneous Response to the main speaker's Initiation. I take backchannelling here to include facial expression and body posture as well as vocalisations. Of course, even in a simultaneous Response, an element of sequence remains. Listeners do not backchannel before they have begun to listen.
3. I have punctuated the interactions conventionally to maximise immediate understanding; obviously there is an element of interpretation in any such punctuation.
4. Much of the data quoted in this paper was collected by third year students on the B.A. Hons in English Language and Literature at the University of Birmingham. I am very grateful to them for both the data and the stimulation and ideas they have provided over the years.

## References

Berry, Margaret (1981) 'Systemic linguistics and discourse analysis: a multilayered approach to exchange structure' in Coulthard, Malcolm & Montgomery, Martin (eds) (1981) *Studies in Discourse Analysis*, London: Routledge & Kegan Paul, pp 120-45.

Brazil, David (1985) *The Communicative Value of Intonation*, Discourse Analysis Monographs No 8, Birmingham: ELR, University of Birmingham

Burton, Deirdre (1980) *Dialogue and Discourse*, London: Routledge & Kegan Paul

Coulthard, Malcolm (ed) (1992a) *Advances in Spoken Discourse Analysis*, London: Routledge

Coulthard, Malcolm (1992b) 'The significance of intonation in discourse' in Coulthard (ed) (1992a), pp 35-50

Coulthard, Malcolm & Brazil, David (1979) *Exchange Structure*, Discourse Analysis Monographs No 5, Birmingham: ELR, University of Birmingham

Daneš, Frantisek (1974) 'Functional sentence perspective and the organisation of the text' in Daneš, F (ed) *Papers on Functional Sentence Perspective*, Prague: Academia, pp 106-28.

Duncan, Starkey, Jr, & Fiske, Donald W (1977) *Face to Face Interaction: Research, Methods and Theory*, Hillsdale, N.Y.: Erlbaum

Fawcett, Robin P; van der Mije, Anita & von Wissen, Carla (1988) 'Towards a systemic flowchart model for discourse structure' in Fawcett, R P & Young, D (eds)

Fawcett, Robin P & Young, David (eds) (1988) *New Developments in Systemic Linguistics*. Vol 2. *Theory and Application*, London: Frances Pinter

Francis, Gill & Hunston, Susan (1992) 'Analysing everyday conversation' in Coulthard (ed) (1992a), pp 123-161, reprinted with modifications from Coulthard, Malcolm (ed) (1987)

*Discussing Discourse*, ELR Monographs No 14, Birmingham: ELR, University of Birmingham, pp 107-48

Grice, H Paul (1975) 'Logic and conversation' in Cole, P & Morgan, J L (eds) *Syntax and Semantics*: Vol. 3 *Speech Acts*, N.Y.:Academic Press, pp 51-8

Halliday, M A K (1961) 'Categories of the theory of grammar' *Word* 17.3.241-92

Halliday, M A K (1985) *An Introduction to Functional Grammar*, London: Edward Arnold

Halliday, M A K & Hasan, Ruqaiya (1976) *Cohesion in English*, London: Longman

Hoey, Michael (1983) *On the Surface of Discourse*, London: George Allen & Unwin, reissued in Reprints in Systemic Linguistics series, University of Nottingham.

Hoey, Michael (1985) 'The paragraph boundary as a maker of relations between the parts of a discourse' *M.A.L.S. Journal*: New Series No 10. 96-107.

Hoey, Michael (1991) 'Some properties of spoken discourses' in Bowers, R & Brumfit, C (eds) *Applied Linguistics and English Language Teaching*, London: Modern English Publications in association with The British Council, pp 65-84.

Lau Hieng Hiong (1992) 'Nominalized Packaging in Scientific Journal Discourse' Unpublished Ph.D thesis, University of Birmingham.

Longacre, Robert E (1976) *An Anatomy of Speech Notions*, Lisse: Peter de Ridder Press

Nesbitt, C & Plum, G (1988) 'Probabilities in a systemic-functional grammar: the clause complex in English' in Fawcett & Young (eds), pp 1-38

Pike, Kenneth L. (1967) *Language in Relation to a Unified Theory of The Structure of Human Behaviour* (1954-9), 2nd rev. edn. The Hague: Mouton

Pike, Kenneth L & Pike, Evelyn G (1977) *Grammatical Analysis*, Dallas Texas: SIL Publications/The University of Texas at Arlington

Ravelli, Louise (1991) 'Language from a Dynamic Perspective: Models in General and Grammar in Particular', Unpublished Ph.D thesis, University of Birmingham

Sinclair, John McH (1972) *A Course in Spoken English: Grammar*, London: Oxford University Press

Sinclair, John McH (1992) 'Priorities in discourse analysis' in Coulthard (ed) (1992), pp 79-88

Sinclair, John McH & Coulthard, Malcolm (1975) *Towards an Analysis of Discourse: The English Used by Teachers and Pupils*, London: Oxford University Press

Sinclair, J McH; Forsyth, I M; Coulthard, R M & Ashby, M (1972) *The English Used by Teachers and Pupils*, final report to SSRC.

Stubbs, Michael (1983) *Discourse Analysis: The Sociolinguistc Analysis of Natural Language*, Oxford: Basil Blackwell

Tannen, Deborah (1989) *Talking Voices: Repetition, Dialogue and Imagery in Conversational Discourse*, Cambridge: Cambridge University Press

Ventola, Eija (1987) *The Structure of Social Interaction: A Systemic Approach to the Semiotics of Service Encounters*, London: Frances Pinter

Wells, Gordon, McClure, Margaret & Montgomery, Martin (1981) 'Some strategies for sustaining conversation' in Werth, Paul (ed) *Conversation and Discourse: Structure and Interpretation*, London: Croom Helm, pp 73-85

# LANGUAGE AWARENESS AND LANGUAGE LEARNING[1]

*Ronald Carter*
University of Nottingham

## 1. Introduction

The topic of language awareness is particularly relevant at present. There is much discussion of language awareness in relation to language development and of the study and analysis of language by language learners, in contexts of teaching English as a mother-tongue as well as teaching English as a second or foreign language. (Donmall, 1985; Papaefthymiou-Lytra, 1987; Hawkins, 1987; Sinclair, 1985). The issues force us to consider how we describe English as a language, the relationships between language study and language teaching, how different language and 'Englishes' can be compared and contrasted and how language intersects with culture and ideology.

M.A.K. Halliday (1987) has suggested that a three-part structure is needed for discussions of language learning:

learning language
learning through language
learning about language

This paper will concentrate on the third part. Students' learning *about* language has declined in recent years. It will be argued that there should be exploration of the pedagogic possibilities of learning about language for the following main reasons. It is valuable in its own right and has an 'educative' potential in the broadest sense of the word. It can enhance learning 'through the language' about the cultures and ideologies which inform the target language and its uses. Finally, it can assist processes of learning a language in so far as knowing about a language is part of knowing a language. This last reason is, however, more controversial, particularly in the era of communicative language teaching.

Discussions of language awareness within the second language acquisition literature focus mainly on grammar. Grammar is likewise prominent in debates within mother-tongue English teaching concerning the place of 'knowledge about language' in the National Curriculum in England and Wales (see Carter, 1990). It is important, however, that discussion also extends to other levels and areas of language. Examples will be drawn here from the domains of vocabulary and idiomaticity, literary language and the interfaces between language and culture and language and ideology. These areas are chosen deliberately because they are normally associated with more advanced stages of language learning and usually

appear, if they appear at all, at the end of a language course. If the relationship between language awareness and language learning is to be explored, such presuppositions need to be challenged. Without drawing undue analogies between processes of first and second language acquisition, we should note that literary uses of language (verbal play with rhymes, puns, nonsense words, jokes, for example) are encountered by young children from the earliest stages. Similarly, idioms, metaphors and cultural embeddings in language are not exceptional but are a natural and pervasive feature of all languages.

## 2. Language Awareness in Action

### *2.1 Language, literature and a small 'l'.*

A useful starting point is with the kinds of 'transparent' language with which we are surrounded daily but which we normally take for granted and do not normally interrogate. Take, for example, the phrase:

Life in the fast lane

and ask yourself in what contexts of use we should expect to encounter such language and what kinds of associations are suggested by it. The phrase is the 'headline' to an advertisement for mens' toiletries (for TURBO After-shave, in fact). Knowing the meaning of these words involves knowing their associations with fast cars, high living and dangerous pursuits as well as with men who are in the fast lane – not only literally in the motorway fast lane but, metaphorically, men whose careers are accelerating due to high levels of performance. Interrogating the phrase further reveals its gender-specificness. It is hard to imagine a perfume called Turbo and, correspondingly, difficult to associate what is fast, powerful and dangerous with women. The gender-directedness of many advertisements can be a key topic in language awareness for language choice is crucial in the mediation of such messages. Advertising language is also a rich source for exploring creative play with words and the social and ideological worlds such play invokes.

Jokes are also inherently creative in the patterns and associations generated by language; they are also significant cultural products as they reveal much of the societies and cultures which shape them and their functions. Jokes are generically diverse and range from straightforward verbal punning, for example:

Q. What is black and white and red all over?
A. A newspaper.

to jokes which allude to or reproduce specific sets of socio-cultural assumptions. For example,

British Rail announces: Coffee up 20p a slice

Here the comic equation of *coffee* (a liquid) with the word *slice* (normally applied to pieces of bread or to cake) together with the sizeable cost of the increase combine to reveal much in public attitudes to British Rail; for example, that British Rail is believed to provide a poor but expensive service; that food and beverages served on British Rail are expensive and of poor quality; that the coffee, in particular, is barely drinkable and is more like bread or cake in its consistency.

The examples here illustrate a basis for awareness of literary and cultural uses of language. In using terms such as 'literary' and 'cultural' these are used with a small

'I' and a small 'c' (see McRae, 1991). In other words, creativity and cultural embedding are not the exclusive preserve of canonical texts but are pervasive throughout the most everyday uses of language (see also, C. & M. Alptekin, 1990; Prodromou, 1990).

### *2.2 Idioms, key words and cultures*

Language awareness can focus both on the patternings of phrases and on individual words. Colours are key words in all languages and, when investigated, often reveal interesting collocational and other patterns. The word *green* in English, for example, refers to a particular point on the colour spectrum but also contracts partnerships of a cultural and idiomatic nature. Thus green is equated with positive actions ('green for go'/'give somebody the green light'); with youth and inexperience ('green'/'greenhorn'); with feelings of jealousy ('green with envy'); with the environment ('green' issues/'the green party'/the 'greens') and with horticultural growth ('greens'/'green fingers'/'greenhouse'). Some of these combinations are transparent and easy to understand; others are more arbitrary; others are clearly culturally significant. There are several ways in which more conscious awareness of such patterns might be stimulated but much can be learned 'about' the word by analysing how such patterns involving *green* (and indeed other colours) are translated into other languages (see McCarthy, 1990, 1991 for further related activities).

### *2.3 Language and ideology*

The relationship between language and ideology can also be productive for generating an important dimension to language awareness. Grammar can play a significant part in such a relationship. Take, for example, the following sentences which are headlines from daily newspapers – each representing a slightly different point of view on events the previous day in the township of Soweto in Southern Africa.

A. Police shoot rioting blacks
B. Rioting blacks shot by police
C. Rioting blacks shot.

In sentence A the police is the clear subject of the sentence and rioting blacks is the clear object of the sentence. The verb *shoot* also clearly links the subject and the object. It is in the active voice and it is a transitive verb. Such specialist terms as 'voice' and 'transitivity' are economic and precise ways of referring to this important verb but much more significant is the function of the verb *shoot*: it clearly makes the police responsible for the shooting; agency is unambiguously assigned to them.

In sentence B the police are still the main agents in the sentence, that is, they are still responsible for the shooting; but the positioning of the noun *the police* is different from sentence A. Here it is relegated to the end of the sentence. Correspondingly, *rioting blacks* is fronted and as a result gets emphasised; items placed at the front of clauses in English receive emphasis. One effect of these choices of grammatical structure is to make it seem as if the rioting blacks are primarily responsible for their own shooting.

In sentence C the agent is deleted entirely. Like sentence B it is in the passive voice. The passive voice allows writers a choice; either the agent can be left in (as with 'by police' in sentence B), or it can, as here, be deleted entirely. In the

process, of course, sentence C presents an interpretation of events in which the police are not involved in the action at all.

The activity should help to reinforce the point that it is important not simply to look through language to the content of the message but rather to see through language to the ways in which messages are mediated or shaped, very often in the interests of preserving or of reinforcing ideologies. Grammar plays its part in such mediation, though it is vital that such a focus on grammar is not on the forms for their own sake but on the functions they encode.

Sentence A is taken from *The Morning Star*; sentence B is taken from *The Guardian*; and sentence C is taken from *The Daily Telegraph*.

A general language awareness involves at least:

a) awareness of some of the properties of language; its creativity and playfulness; its double meanings.

b) awareness of the embedding of language within culture. Learning to read the language is learning about the cultural properties of the language. Idioms and metaphors, in particular, reveal a lot about the culture.

c) a greater self-consciousness about the forms of the language we use. We need to recognise that the relations between the forms and meanings of a language are sometimes arbitrary (witness the above example of *green*) but that language is a system and that it is for the most part very systematically patterned.

d) awareness of the close relationship between language and ideology. It involves 'seeing through language' in other words (see Carter and Nash, 1990).

## 3. Language Awareness and Teaching Languages

What are the theoretical implications of this kind of analysis? What is its relevance to language teaching? The first observation is that native speakers of a language and advanced learners of a second or foreign language react to such bits of language mainly unconsciously and unreflectingly. In normal circumstances of communication – where there is successful uptake – most language users do not analyse language in this way. There is no need to. Indeed, some would argue that, for example, analyzing jokes can kill them, can destroy their effectiveness. By analogy, therefore, many language teachers would argue that we should not be encouraging our learners to analyse the target language. It is often said that too much self-consciousness can restrict opportunities for language acquisition and inhibit the learner. They point to the prevalence of communicative language teaching practices which are designed not to help learners to analyse the language but to experience it *in use*. Communicative methods do not, indeed should not, cultivate reflectiveness on language.

Opponents of language awareness also argue that in order to analyse language a considerable range of metalanguage is needed – that is, you need language to talk about language. (Notice in this connection the terms employed so far such as active, passive, agent-deletion, opaque/transparent idioms etc.) It is said very forcefully that language learners have a long journey to take. We should not make it more difficult for them by giving them extra luggage to carry. If language awareness is extra luggage, then learning a metalanguage is definitely excess baggage.

Of course, the rejection of analysis by communicative language teaching theorists is part of a reaction against structural methods in language teaching. For example, grammar translation methods involved a lot of conscious metalingual naming of grammatical parts. There was no corresponding attention to helping learners use the language fluently in authentic contexts. Audio-lingual methods did not draw attention to language structure as explicitly as grammar translation methods. But audio-lingual methods are based on an isolation of language structure – a declarative knowledge which teachers seek to convert into procedural knowledge by pattern practice and the use of drills.

Communicative language teaching, influenced by theories such as those of Stephen Krashen, states that languages are acquired rather than learned. Indeed, language learning and language acquisition can be seen to be opposed. Language learning is largely a form-focussed activity. It focuses on the structures of the language in an explicit way. It focuses on accuracy. Language is taught as if it were a product, a static, machine-like entity. Such learning results in learners knowing *that* rather than knowing *how* – knowing that, for example, certain rules obtain in particular uses of language. Knowing *that* is conscious knowledge. It is language awareness. Such knowledge may act as a monitor or editor of language use but is not and cannot be equivalent to language use.

For Krashen and for others greater emphasis should be placed on classroom language activity which is meaning-focussed and with a focus on fluency. Meaning-focussed activity exposes learners to and immerses them in language. It helps generate implicit knowledge of the language. Language is taught as if it were a process, an organic, dynamic entity. Such processes of learning are, it is said, much more likely to result in knowing *how* rather than knowing *that* – that is, knowing *how* to use language fluently, unselfconsciously and without inhibition. It is an intuitive implicit knowledge *of* the language, not an explicit knowledge *about* the language.

It can be seen, therefore, that language awareness is resisted from a number of theoretical and practical viewpoints. Indeed, familiar contrasts and dualisms impregnate our discourse:

| | | |
|---|---|---|
| implicit | v. | explicit |
| meaning-focussed | v. | form-focussed |
| declarative | v. | procedural |
| conscious | v. | unconscious |
| knowing how | v. | knowing that |
| product | v. | process |
| language as static | v. | language as dynamic |
| knowledge of | v. | knowledge about |
| accuracy | v. | fluency |
| language learning | v. | language acquisition |

Such familiar dualisms in discussions of language lead to what has been termed the pendulum theory of language teaching. Once a pendulum has swung one way, then it should swing at least as far the other way. In this respect communicative methodology is at one end of the swing of the pendulum; structural grammar translation and audiolingual methods are at the other end of the swing of the pendulum – a swing which, in fact, takes us in a completely opposite direction. Until recently

and during the 70s and 80s the pendulum has swung firmly away from and against language awareness.

Instead of adopting a dualist perspective, I would like to explore the possibilities of integration of these seemingly opposed theories of language learning. I believe that the development of a form of language awareness can serve such integration. Such integration can bring together conscious and unconscious approaches to L2 language development. This perspective is endorsed by Henry Widdowson in a chapter in his recent book *Aspects of Language Teaching*:

> ...it seems on the face of it to be likely that with some learners a conscious awareness of how language works and the subjection of their experience to analysis would suit their cognitive style, increase motivation by giving added point to their activities, and so enhance learning. It would enable them to make comparisons between the language they are learning and their own language, and engage in the kind of rational enquiry which is encouraged in other subjects on the curriculum.
>
> Widdowson (1990);

and by Ellen Bialystok in a paper published in 1982:

> In unanalysed representations of language, only the meanings are coded; in analysed representations, both the meanings and the relationships between the forms and those meanings are coded. Such analysed representations permit the learner to manipulate those form-meaning relationships to create particular structured uses of language. While conversations may proceed perfectly well from unanalysed representations .... other uses of a language involved in reading, writing, lecturing and explaining depend on greater analysis in linguistic structure.
>
> Bialystok (1982)

The quotations illustrate that explicit and implicit knowledge need not be in opposition and that under certain conditions explicit knowledge can facilitate acquisition. An important contribution to this debate is William Rutherford's book *Second Language Grammar*. Rutherford makes a key point when he writes:

> ... whatever it is that is raised to consciousness is not to be looked upon as an artifact or object of study to be committed to memory by the learner.... what is raised to consciousness is not the grammatical product but aspects of the grammatical process.... C-R (consciousness raising) activity must strive for consistency with this principle.
>
> Rutherford (1987)

## 4. Integrating Language Awareness: The Contrastive Principle

If we accept that consciousness-raising or language awareness can be more extensively introduced into the language classroom, then how is this best achieved? If we accept that it is likely to be more successful if it is not seen as a separate classroom activity (a 20 minute slot on language awareness on Thursday at 2.30) but rather integrated into the ongoing process of language learning, then how is this best achieved? If we accept that learning about language can inform

not just language learning but learning in the broadest sense of the word, then how is this best achieved?

This brings us now to the second main question posed at the beginning of section 3. What is the *relevance* of language awareness to language teaching and what does this mean in the context of the language classroom?

Explorations in this area in both mother-tongue and foreign language teaching underline the value of adopting a *contrastive principle*. A contrastive principle states that we are more likely to see things perceptively, creatively and with understanding if things are viewed not in isolation but set alongside each other, compared and contrasted. There are innumerable opportunities within the system of a language for contrasts to be generated.

To activate this principle three broad parameters of language awareness can be posited:

a parameter of *form*
a parameter of *function*
a parameter of *socio-cultural meaning*

### *4.1 Form*

Activities within the form parameter involve a systematic focus on more formalistic aspects of language. Examples might include strategies which draw attention to the *-ed* ending in seventy per cent of English past tense verbs; the frequency of plural in *s*; the phenomenon of *th* (ð and θ) in English phonology; the contrast between count and non-count nouns in English (see Ellis, 1989). Control of such forms is important for accurate use of the language. Such parameters of form can be usefully foregrounded by comparisons between the target language and the learner's language and/or interlanguage. Numerous activities exist or could be developed which might foster enhanced awareness of such formal properties of language.

There is always a certain arbitrariness both to forms of language and in the relations between form and meaning. Lexical collocations are a good example of this. Thus, you can have:

| | |
|---|---|
| | a strong argument |
| | a powerful argument |
| and | |
| | strong tea |
| but *not* | powerful tea |
| You can have | |
| | dry ground |
| | wet ground |
| and | dry bread or toast |
| but *not* | wet bread or toast |

These kinds of lexical gaps can be best exploited within activities which both highlight the contrasts and gaps internal to the target language and which bring into conscious awareness one or more of the related or contrasting patterns within the learner's language. An important component of the contrastive principle is the need to draw attention both to what is there and what is not there in and across languages. We should also note research which underlines the importance of learning

words within contrastive patterns and sets. Such processes facilitate the kind of cognitive depth which is central to memorization of words.

*4.2 Function*

Activities within the parameter of function are designed to raise awareness of what language does, particularly in communicative contexts. Such awareness involves looking at the relationship between language and contexts of use. There are several classroom possibilities here:

i) comparisons between spoken and written texts (especially spoken and written versions of the same content or theme)
ii) comparisons, preferably in relation to a common content, between different stages in the history of English or between different international Englishes
iii) comparisons between different *translations* of the same stretch of language (cf. Duff, 1990)
iv) comparisons between contrasting styles – designed for different purposes or functions (e.g. real language v. textbook language; scripted v. unscripted talk; real and made up examples in dictionaries. COBUILD data is especially valuable here.) (See also Willis, 1990)

More specifically, it can be productive to generate awareness of the functions of words and phrases in texts, especially conversations. Studying the ways words are used to close down conversations can set up perceptions of how words can have different meanings and functions in different contexts (often underlining in the process the arbitrariness of the form/meaning relationship). It can also be fun to work out how words like *right, OK then, good* or phrases like *I'll let you be going then* or *This call must be costing you a lot of money* can be used to signal a desire to finish a telephone call. Activities of this kind illustrate the importance of understanding how closely language function and situation are intertwined (see recent articles by Bardovi-Harlig *et al*, 1991 and Holborrow, 1991).

*4.3 Socio-cultural meaning*

Awareness within the parameter of socio-cultural meaning is also best achieved by invoking the contrastive principle. Examples here might include activities which generate awareness of language cross-culturally. As we have seen, differences in newspaper headlines are an obvious starting point but within different newspapers and magazines the language of horoscopes, agony aunt letters or wanted columns involve different cultural and social assumptions. Indeed the *absence* of such items within a particular English language newspaper in different parts of the world or within the newspapers of the learners' culture as a whole could raise numerous points for contrastive cultural and ideological analysis.

The existence of three parameters should not imply that they are wholly separate or discrete. The Gulf War has, for example, made us aware of the finite nature of oil resources. Yet *oil* is a non-count noun (like *water, air, petrol*). The form contains an implicit perception that such a resource is limitless and unbounded. The next century may determine a change in the grammar. We may have to talk of units of *oil* or *petrol*, an *oil* or a *petrol* or *oils* and *petrols*. The example illustrates the interconnection between formal, functional and socio-cultural parameters, with grammar and ideology closely embedded within each other. Thus teaching about

the system of count/uncount nouns as forms can also be an opportunity to integrate cultural and language awareness. All the three parameters provide rich opportunities for cross-lingual and inter-lingual comparisons. Such opportunities should, it must be said, arise within the context of meaningful classroom activity; otherwise such language awareness can appear to both teacher and pupil as something a little too mechanistic and contrived.

## 5. A Reflective Language Learner: A further example

One argument not so far mentioned is the case for an increased learner autonomy which goes with increased language awareness. Consciousness-raising in the area of language form and structure is closely connected with the movement in recent years to give to learners greater control over their own learning. One particular domain here is learner training, the notion of learning to learn English promoted, very successfully in my view, in the work of Gail Ellis and Barbara Sinclair. The aim of the teaching materials developed by Ellis and Sinclair (1989) is to promote greater awareness on the part of learners of the learning strategies which they use. Such greater consciousness will, it is argued, help make such learners more reflective, flexible and adaptable. *A more reflective language learner is a more effective language learner.*

Let me bring this paper to a conclusion with an example of language awareness in relation to a literary text and in the process develop a case for an increased use of literature in the language classroom. Here is the first stanza from 'yes is a pleasant country' by the American poet e e Cummings

> yes is a pleasant country
> if's wintry
> (my lovely)
> let's open the year.
>
> (from *Complete Poems 1913-1962*, published by HarperCollins)

The poem uses very simple language. But the poem is ungrammatical and it is also semantically deviant. We don't open years; subordinators do not normally appear in subject position. *Yes* cannot be a country, and so on. But I have watched with fascination how groups of students in many parts of the world, sometimes discussing in English or in their mother tongue – according to level – begin to unpick its meanings, begin to interpret it, begin to make it make sense.

Most groups end up with a reading which takes its cue from the brackets (*my lovely*). It is read as a love poem or an interchange between lovers. The speaker is trying to persuade his or her lover to say yes, to be affirmative and positive – to make 'yes' not deviant but normal. To keep saying *if* imposes conditions (*if* is a conditional); it makes the response cold and unpleasant. The speaker is appealing for a new start, for a new beginning to a new year. Let's move (metaphorically) from a cold to a warm country. Saying *yes* is warm; saying *if* is cold.

The discussion of this kind of text helps generate language awareness; it shows that rules can be broken for creative purposes. It also helps to foster interpretive skills and to encourage reading between the lines of what is said. It can help teach the confidence to make sense of language input which is not always – in real communicative contexts – neat, clear and immediately comprehensible (see Carter and Long, 1991).

## 6. Conclusion

I have argued in this paper for the following:

1. Learning a language involves *understanding* something of that language. It is unlikely that such understanding can be developed by naturalistic exposure. It has to be quite explicitly taught.
2. Teaching can and should build on existing competences. For example, all learners have an inbuilt *literary* competence. But it has to be developed.
3. Learning about a language also involves understanding something of the culture within which the language is embedded. This involves aesthetic understanding, appreciating the creative play and invention of language use. Knowing a language involves appreciating how and why its rules can be broken or creatively manipulated. It involves appreciating jokes and ironies, responding to puns, and seeing through language to the points of view and ideologies which it can reveal as well as conceal (see also Candlin, 1989).
4. Such language awareness assists in the development of interpretative and inferential skills. Indeed it has to. It is impossible to teach in detail about the literature, the culture or ideologies of the societies which use the target language. There is neither time nor curricular space to allow this. What can be taught is the procedural ability, the ability to learn how to learn such things, the capacity for interpretation and inference in and through language.
5. In this way, language learning and teaching become indistinguishable from language *education* in the broadest sense of the word. Learners are better learners if they are able to analyse what they are doing and *why* they are doing it. Language teaching has for too long been seen as training in the instrumental functions and purposes of the language – and for too long there has been a strong anti-intellectualism associated with communicative language teaching. A learner, educated in the use of the language, is aware of the language as a cultural artefact. She/he is a student/analyst of the language as well as a user. Such awareness can be stimulated at all stages in the language development process and courses should make greater provision for developing such awareness – from the earliest stages to the most advanced levels.

In England the LINC Project which was based in the English Department at Nottingham University aimed to devise materials which would help teachers to implement the new National Curriculum for English in England and Wales. Part of the brief was to develop a range of tasks and procedures for helping pupils acquire greater knowledge about language or language awareness.

Teachers have responded to this work with great enthusiasm. They are particularly interested in the relationship between KAL and using the language effectively and interesting evidence concerning this relationship (which would need to form the subject of a separate paper) is being gathered. Many teachers do, however, point out that greater language awareness confers greater power on pupils. The ability to analyse language should not be the exclusive preserve of the teacher. It is important that teachers share that power with learners.

Work on language awareness in English as a mother-tongue is commensurate with current developments in EFL/ESL in learner autonomy, task-based learning,

student-centred language development, the relationship between teaching the language and teaching the culture and with the pedagogic implications of the shifting roles of teachers and learners in the classroom.

Let me conclude by conceding that much in this paper is speculative and that investigations into the relationship between language awareness and language teaching are still at an early stage. In particular, the integration of language *learning*, learning *through* language and learning *about* language is very much a hypothesis. But I hope it is a hypothesis which can be more extensively tested and developed in the 1990s. Given that the question of language awareness or knowledge about language is being explored both in English as a mother-tongue teaching – especially in connection with the new National Curriculum – and in EFL and foreign language teaching, I would hope that it might lead to greater integration between fields normally kept separate but which obviously have much in common, actually and potentially. This would be an integration which, I suspect, John Sinclair, as someone who has always had a foot firmly in both camps, would warmly support.

## Note

1. Parts of sections 3 and 4 of this paper appear in McCarthy, M. and Carter, R. *Language as Discourse: Perspectives for Language Teaching* (Longman, Harlow, 1993) and are reprinted here by kind permission of Longman Group.

## References and Select Bibliography

Alptekin, C. and M. (1990) 'The questions of culture: EFL teaching in non-English-speaking countries' in Rossner, R. and Bolitho, R. (eds.) *Currents of Change in English Language Teaching* (OUP, Oxford) pp. 21-26.

Bardovi-Harlig, K. et al (1991) 'Developing pragmatic awareness: closing the conversation', *ELT Journal* 45, 1, pp. 4-15.

Bialystok, E. (1982) 'On the relationship between knowing and using linguistic forms' *Applied Linguistics*, 3, pp. 181-206.

Candlin, C. N. (1989) 'Language, culture and curriculum' in Candin, C.N. and McNamara, T.F. (eds.) *Language, Learning and Community* (NCELTR, Sydney) pp. 1-24.

Carter, R.A. (ed.) (1990) *Knowledge About Language and the Curriculum: the LINC Reader* (Hodder & Stoughton, Sevenoaks)

Carter, R.A. and Long, M. (1991) *Teaching Literature* (Longman, Harlow)

Carter, R.A. and Nash, W. (1990) *Seeing Through Language: A Guide to Styles of English Writing* (Blackwells, Oxford)

Donmall, G. (ed.) (1985) *Language Awareness* (CILT, London)

Duff, A. (1990) *Translation* (OUP, Oxford)

Ellis, G. and Sinclair, B. (1989) *Learning How To Learn* (CUP, Cambridge)

Ellis, R. (1989) *Instructed Second Language Acquisition.* (Blackwells, Oxford) esp. ch. 7.

Halliday, M.A.K. (1987) 'Some Basic Concepts in Educational Linguistics', in Bickley, V. (ed.) *Languages in Education in a Bi-lingual or Multi-lingual Setting*. Hong Kong: ILE, pp. 5-17.

Hawkins, E. (1987) *The Awareness of Language* (rev. ed. CUP, Cambridge)

Holborrow, M. (1991) 'Linking language and situation: a course for advanced learners', *ELT Journal* 45,1, pp. 24-33.

McCarthy, M. (1990) *Vocabulary* (OUP, Oxford)

McCarthy, M. (1991) *Discourse Analysis for Language Teachers* (CUP, Cambridge)
McRae, J. (1991) *Literature with a small 'l'* (Macmillan, London)
Papaefthymiou-Lytra, S. (1987) *Language, Language Awareness and Foreign Language Learning* (Univ. of Athens Press, Athens)
Prodromou, L. (1990) 'English as cultural action' in Rossner, R. and Bolitho, R. (eds.) *Currents of Change in English Language Teaching* (OUP, Oxford) pp. 27-39.
Rutherford, W. (1987) *Second Language Grammar* (Longman, Harlow) esp. chs. 2 and 5.
Sinclair, J.M. (1985) 'Language awareness in six easy lessons' in Donmall, G. (ed), *Language Awareness* (CILT, London) pp. 33-37.
Widdowson, H.G. (1990) *Aspects of Language Teaching* (OUP, Oxford) esp. chs. 6 and 10
Willis, D. (1990) *The Lexical Syllabus* (HarperCollins, London).

# CONQUEST OF PARADISE – LANGUAGE PLANNING IN NEW ZEALAND[1]

*Robert B. Kaplan*
University of Southern California

## 1. Introduction

The title of this paper is intended to evoke the recent American film about the discovery of the "New World" by Christopher Columbus, produced in 1992 to commemorate the 500th anniversary of that event. The film is visually attractive, but a bit thin on plot. Despite the best efforts of the film's director and its principal actors, Columbus remains a fairly unattractive figure, and the genocidal after-effects of the discovery are not much mitigated. One is reminded of Mark Twain's well-known comment to the effect that the fact that Columbus had discovered America was not at all remarkable; had he missed it, that would have been remarkable. Although not as early, and not commemorated in epic film, the discovery of New Zealand and the after-effects of that discovery are not dissimilar. It is important to remember that the conquered paradise was in the eye of the beholder. To Columbus and his shipmates, the New World was a paradise which they exploited for their benefit; to the indigenous people, the New World was not new, had been theirs, and was a paradise lost to the foreign conquerors; so in New Zealand, Englishmen found a paradise to exploit, and the Maori people lost a paradise to exploitation.

## 2. History

The "discovery" of New Zealand by Europeans occurred some 150 years after Columbus "discovered America" (Abel Tasman, a Dutch sea-captain in the employ of the Dutch East India Company, arrived in 1642) and with a rather different cast of characters (Captain James Cook arrived in 1769, and the New Zealand Company sent the first English settlers in 1839), but the after-effects have been remarkably parallel (despite the signing of the Treaty of Waitangi between the Maori chiefs and British Lieutenant Governor William Hobson in 1840). Polynesian navigators had, of course, come upon New Zealand some time before Europeans did (about 200 years before Columbus stumbled into the New World), had settled it, had conquered the few small Moriori tribes which were already in possession, and had lived in it with moderate comfort and success. It

is not necessary to romanticize the Maori period in New Zealand's history; the Maori fought each other, and their existence was not, at least in some senses, paradisical, but the land was beautiful, the food supply was adequate, there were no competing predators, the climate was not severe, and the isolation was protective of their way of life. It was Englishmen who thought they were coming to an earthly paradise; for Maori people it was just home.

Despite the Treaty of Waitangi, despite the promises to protect the Maori language and culture, over the ensuing hundred years Maori language and culture were ravaged (Benton 1981). English became, *de facto*, the official language of New Zealand – indeed, for all practical purposes the only language (Kāretu 1991). Over the past half century, however, New Zealand has not only allowed but encouraged immigration, and Dutch refugees from World War II, Polynesians from New Zealand's United Nations designated protectorates in the South Pacific, other Europeans (i.e., Greeks, Serbo-Croatians, Italians, etc.), a smattering of Turks, Chinese, and Gujaratis, and, most recently, a fairly steady trickle of Indo-Chinese have settled in New Zealand (Hirsh 1987). The arrival of this olio of migrants, together with the recent emergence of a more militant attitude among the relatively few surviving Maori, have created a very complex language situation (Waite 1992). The situation has been further complicated by some redefinition of educational needs and by the quite recent recognition of the needs of several special groups. It shall be my purpose to describe the current language situation as I see it, and to comment on the current (excessively belated) efforts to create a National Languages Policy which will in principle take account of the real linguistic diversity of contemporary New Zealand and will seek to preserve linguistic diversity but at the same time speak to the needs of some form of common communication (Kaplan 1992).

## 3. The New Zealand Language Situation

It is difficult to be very specific about the language situation in New Zealand because no linguistic data bases have ever been maintained. Some information is available with respect to languages taught in the educational system, but even this data is incomplete for some languages (AGB/McNair 1992, Benton 1991b, Hawley 1988, G. D. Kennedy 1982, Levett and Adams 1987). Notional information would suggest, however, that a large number of languages is actually spoken (or used) in New Zealand. The languages can be broken down into five broad categories: Maori, other Polynesian languages, a broad category which can only be designated "other languages," the small set of international languages, and English serving a variety of functions.

### *3.1. Maori*

Maori must be treated separately because of its special political status—ironically, it is the only *de jure* official language of New Zealand under the terms of the Maori Language Act (Government of New Zealand 1987). There are complex questions about Maori; a debate continues among Maori speakers whether Maori should be a written language (though a written form exists and has existed for some time). There is no standard spoken Maori, since various Iwi (tribal groups) espouse various versions of the language; there is no agreement whether a standard

spoken variety of Maori ought to exist, even for educational and political purposes. It is the case that the number of first-language speakers of Maori has been declining steadily, though in recent years, thanks largely to the efforts of the Maori people themselves, a relatively large number of bilingual second-language speakers of Maori has emerged (Fishman 1991, Grace 1991, Hirsh 1990, Ministry of Education 1989b, Spolsky 1987). There are unresolved questions about the role of Maori in New Zealand society; e.g., though it is used in the bilingual designation of most government agencies ought it also to be represented in New Zealand's passport, in New Zealand's currency, in New Zealand's postage? Should it be required in the civil service? Should tertiary students be able to submit examination papers in Maori (in disciplines other than Maori studies)? What should its role be in the judicial system (Bates 1991)? Who should learn Maori; when should it be taught and learned, and who should be responsible for paying the costs involved in teaching it (Dunn, Pole, and Rouse 1992, Sexton 1990, Thomson 1991, Visser and Bennie 1991)? What is the teacher pool, and is it likely that there will ever be a sufficient pool of teachers? Is Maori becoming essentially a ritual language; that is, is it constantly losing registers to English, leaving available only a limited set of ritual registers? What relationship between language and religion can be devised? (Many Maori speakers (whether first – or second – language speakers) have adopted Christianity, and much of the contemporary use of Maori in public situations carries some Christian religious message, but it is clear Christianity is not entirely compatible with Maori values and vice versa, and there is an important question of the interrelationship among *Maoritanga*, Christianity, and English.) Clearly, the future of *Te Reo Maori* (Maori language) is neither assured nor unclouded.

### *3.2. Other Polynesian Languages*

The other major Polynesian languages spoken in New Zealand are **Cook Island Maori, Niuean, Samoan, Tokelauan**, and **Tongan**. The populations of speakers are, except for Samoan, relatively small (even the Samoan population is small in absolute numbers and as a percentage of the New Zealand population, though it is considerably larger than the numbers of speakers of other Polynesian languages); but then the populations of the islands from which these languages originate are also small, and it appears that the survival of these languages (except perhaps for Samoan) is not at all assured even in their home territories, let alone in New Zealand. The same sorts of questions raised in conjunction with Maori language and in conjunction with Christianity apply also in the situations of each of the other Polynesian languages, and the questions are even more urgent because these languages, unlike Maori, are not protected by treaty and legislation (Iosia 1992). Furthermore, there is a long-term tendency among Pakeha (European) New Zealanders to view the situation of the several Polynesian languages as monolithic and to fail to recognize that each of these languages (and each of the language communities) has its own distinct problems and its own special needs.

### *3.3. Other Languages*

#### 3.3.1. What languages are in the mix?

The broad category of "other languages" can be subdivided into four sub-categories:

(i) other Indo-European languages like Dutch, Modern Greek, Serbo-Croatian, and Spanish;

(ii) non-Indo-European languages like Turkish and Chinese (largely Cantonese, but also Mandarin), as well as Modern Hebrew, South Asian languages (largely Gujarati but also Tamil), and Southeast Asian languages (e.g., Cambodian, Hmong, Lao, and Vietnamese);

(iii) classical languages like Anglo-Saxon, Classical Arabic, Classical Greek, Classical Hebrew, Latin, Old Church Slavonic, Old High German, Old Icelandic, and Sanskrit; and

(iv) languages of special communities, like New Zealand Sign Language used by the deaf community (Brian, Dugdale and Logan 1992)

Clearly, category iii above is different from the other categories in the sense that this set of languages have purely academic value while the other categories contain languages which are spoken in contemporary New Zealand. All of these sets deserve attention, though the kind of attention given the classical languages will surely be different from that given the living languages.

### 3.3.2. How are they maintained?

Many of these languages are maintained by their communities of speakers, either through efforts in the family or – in the more affluent communities – through private tutoring and Saturday schools. Some of these languages (especially the classical languages) are taught exclusively in the school system. Some of these languages have become indigenized in New Zealand, others remain largely "foreign" or "academic". To a significant degree, these languages largely are not recognized in the educational system and are not supported through governmental funding. By the same token, standards for achievement, general criteria for teacher training, standards for curriculum, standardized assessment instruments, quality textbooks, and – simply – adequate supplies of reading material do not exist for many of these languages. In sum, these languages – to a large degree – survive despite the efforts of the government (McGregor and Williams 1991). There is no formal "English only" movement in New Zealand probably because there isn't any need for one; the hegemony of English is virtually complete. Yet these languages individually and collectively constitute a matter of some urgency for New Zealand society (Kaplan 1980,1981).

### 3.3.3. Economic Languages

A newer thrust which must be considered in this category is the very recent interest of government in "commercial" languages; the government holds the view that New Zealand's international trading position can be enhanced by developing, in the broad community, facility in the languages of its trading partners – at the moment such languages as Chinese (Mandarin), Indonesian, Japanese, and Korean (Bolger 1992, Panitchpakdi 1992, Smith 1992). So far there appears to have been a failure to recognize that this is ephemeral – that is, that because the international economic situation is fluid, and because trading partners consequently change rather frequently, this list of languages inevitably also changes. For example, Russian and Thai are not currently included in the list, though they are likely to increase in commercial importance over the next decade and, although Chinese,

Indonesian, and Korean are represented, the amount of resources devoted to promulgating them is negligible. The failure to understand the ephemeral character of this list has tended to obscure the time it takes to develop a capability to teach a language, the time it takes to educate any significant number of speakers, and the time it takes for these speakers to penetrate the labour market. The chances are that the need for any particular language may diminish or disappear before the manpower pool for that particular language is adequately developed; that is, language education is likely to be outstripped by changing economic needs. There is also a comparable failure to recognize that it is easier to achieve fluency in, say, Chinese, by training a Chinese New Zealander who already has some facility in that language than it is to train a monolingual English-speaking New Zealander "from scratch". And there is a comparable failure to recognize that the international trading position of a nation depends more on the kinds of products the nation sells, on quality standards, on the ability to deliver in a timely fashion, and on cost than it does on bilingual fluency though it cannot be denied that bilingual fluency plays some yet undefined role in the process. There is little question that one can buy in the international market place in English, but that it requires some proficiency in the languages of other potential buyers to sell to those buyers. However, bilingual fluency in the population may be achieved in many ways – teaching indigenous English speakers to be bilingual (while it is for many reasons highly desirable) is only one of the ways to approach the economic issues. Besides, there is an important sociolinguistic question whether potential buyers would prefer to deal with a Pakeha New Zealander who speaks English and Chinese or with an ethnic Chinese of any origin who speaks Chinese and English. There is evidence that, for example, Japanese tourists in New Zealand prefer Japanese guides who are English-speaking to New Zealand Pakeha guides who speak Japanese, particularly if they speak Japanese haltingly. It is demonstrably the case that English-speaking New Zealanders trained in Japanese are not able to find work utilizing their linguistic skills; employers report that these learners have too narrow a language facility (largely literary)[2] and that they lack relevant training in business (Levett and Adams 1987). In sum, the many issues involved in the economics of language are not well understood and, while the economic motivation for language development is not trivial, it will be important to gather a clearer notion of the economics of language across the region (Coulmas 1992).

### 3.3.4. Special Languages

The deaf community has developed, in New Zealand, a special variety known as New Zealand Sign Language (NZSL). This sign language is distinct, unique to New Zealand, and different from sign languages used in Australia, the United States, and other parts of the world. It is not merely a way of representing English but it is in every sense a natural language with an independent grammar and lexicon. It has developed in the deaf community and has largely been ignored in the hearing community. At the present time, the first dictionary of NZSL is being prepared at Victoria University in Wellington. But there are inadequate resources for the dissemination of NZSL not only within the deaf community but within that segment of the hearing community that interacts with the deaf community on a regular basis.

### *3.4. International languages*

In the various political arrangements emerging out of the international movement which followed World War II, certain language agreements were politically achieved. Chinese, English, French and Russian emerged as the languages of the United Nations by political accord. Because various agencies of the United Nations quickly expanded into the world-wide storage and management of information, these languages carried over into that sphere. A number of complex causes have co-occurred in this context:

- because the United States was the only major industrialized nation to emerge from the War with its educational and industrial infrastructure completely intact,
- because for several decades U.S. scientists both most used existing information systems and most contributed to those systems, thus coming to "own" the systems,
- because the U.S. educational system became a magnet for the youth of the developing world and for the youth of nations rebuilding from the devastation of the War,
- because the development of electronic international information networks exactly corresponded with the emergence and development of the modern computer,
- because the quantity of information available has been expanding geometrically[3],
- therefore, English has achieved an unprecedented hegemony in the world's information storage and retrieval systems, in science and technology, and in other global registers.

Chinese was essentially overlooked because the foundation computers were unable to deal with Chinese characters and because political events in China tended to isolate Chinese scholars. Russian too was restricted in its contribution to the systems because of the position taken by the government of the USSR during the period of the Cold War. The outcome is that something on the order of 85% of all the scientific and technical information available in the world today is either originally written in, or abstracted in, English.

For this complex reticulated set of reasons, essentially three languages have emerged as international languages: English and French, and more recently (as Germany's role has increased in the European economy), German. The hegemony of English in this context continues, however, and the English speaking nations possess an information cartel of absolutely staggering proportions[4] (Baldauf and Jernudd 1983, Kaplan and Medgyes 1992). For most of the time since the end of World War II, the English speaking nations have not consciously exercised the power of that cartel, but in the last dozen years, in the name of national security and the protection of patents and copyrights (the down-side of the process was designated the "technology haemorrhage" by the Reagan administration), the United States has begun to invoke its informational power and to limit the free flow of scientific and technical information.

Be that as it may, the world generally recognizes the existence of English, French and German as international languages. (It is important to note that the sort

of hegemonic international status discussed here has nothing to do with the absolute or relative size of the population speaking the language; if size were the sole criterion, at least Arabic, Mandarin, and Spanish would certainly be candidates for international status). As English, French and German have become increasingly international they have also become decreasingly "national" and decreasingly "culture-bound". The English which is used in the great international information storage and retrieval networks is not an English regularly spoken by any native English speaker. International English is not the property of any cultural grouping, of any society, of any nation; it is genuinely an international *lingua franca*, controlled in part by certain academic-disciplinary discourse communities. To some extent, the same thing is happening to French and German, and the pace of internationalization of these languages is likely to increase as they become the linguistic vehicles of the new united states of Europe[5].

New Zealand recognizes the great importance of these international languages – indeed it depends upon the international status of English – but it does not recognize that standards of correctness in the international domains are not the property of any community of speakers. New Zealand has yet to develop appropriate mechanisms for dealing with this set of international languages *qua* international languages as assets to the society and as curricular subjects in the educational system, though it does deal with them as representatives of particular cultural polities – England, France, and Germany.

### *3.5. English*

While English is unquestionably the *de facto* official language of New Zealand (but not the *de jure* official – or national – language), it is not itself free of problems. There are probably at least five sets of issues that beset the status of English in New Zealand.

#### 3.5.1. Which English?

There is a clear preference among New Zealand speakers of English for the educated British variety. That is not to say that the British variety is consciously taught in schools; rather, it is fairly clear that many New Zealanders have an emotional preference for that variety and consider it a prestige (High) form. There is, however, a New Zealand variety, different from the British standard, or for that matter from the Australian and North American varieties, widely spoken in New Zealand. That variety needs to be described, and New Zealanders need to determine how they feel about it (Bell and Holmes 1990). It is now possible to undertake language description on the basis of massive electronically stored corpora of a language (as is presently being done in at least the COBUILD project at the University of Birmingham, in the independent "British National Corpus" being assembled by a consortium of British publishers, and in the 13-nation "International Corpus of English" (to which New Zealand has contributed through the work of G. Kennedy) being coordinated by S. Greenbaum), and it seems highly desirable for such descriptive work to be undertaken on New Zealand English in New Zealand.

### 3.5.2. English as a Second/Foreign Language

As the number of migrants of various categories has increased in New Zealand, and as greater proportions of those migrants originate in non-English speaking countries, an increasing need for instruction in English as a second or foreign language has developed. Present efforts in that context are inadequate both because they deal with too small a fraction of the New Zealand population at need and because the efforts are too constrained in terms of resources and instructional time (G.D. Kennedy 1988, Ministry of Education 1989a). There has been a culturally-based addiction to a somewhat xenophobic perception that the inability to speak English "properly" is a sign of ignorance and that the appropriate treatment of the problem implicates remediation. No one would claim that the teaching of a foreign language – French, Japanese – to New Zealanders is a remedial activity, but there is a wide-spread belief that speakers of other languages learning English should be treated as "bone-head" remedial learners.

In recent years, New Zealand has become increasingly interested in the export of English-as-a-Foreign-Language instruction in the South Pacific, Australasian region. Such exportation of language instruction is seen to have commercial value; it constitutes an important cash export, bringing new monies into the New Zealand economy (as it has brought new monies into the U.S., Australian, Canadian, and Irish economies, and as it has been a substantial element in the U.K. economy for a good many years). New Zealand's version of the Peace Corps – Volunteer Service Abroad (VSA) – has for some time engaged in the delivery of English language instruction on a not-for-profit basis in the less developed nations of the region. But commercial organizations have been established both to teach English for profit in the countries of the region and to import foreign students for relatively short periods to study English in New Zealand.

New Zealand is a late-comer into this market, and it would be fair to say that the market is not well understood either in the New Zealand educational and commercial sectors or in the government (e.g., witness the rather unpleasant situation arising from the exploitation of Mainland Chinese students by some New Zealand proprietary language schools).

### 3.5.3. Literacy

Although New Zealand enjoys (and has for some time enjoyed) an international reputation for the excellence of its English literacy programs, current economic problems threaten to undermine this high quality (Wagemaker, mimeo). Additionally, despite the excellence of the programs, some segment of the English-speaking and non-English-speaking populations of New Zealand has fallen through the cracks and is not adequately literate. This is not merely a linguistic problem. Literacy is a remarkably flexible commodity. Elite cohorts of any community to a significant degree define the nature of literacy at any given moment. As population pressure increases on the elite cohorts – as greater numbers of individuals wish to enter the ranks of the elite than that elite can comfortably accommodate – those cohorts have the ability to "raise the ante"; that is, those cohorts can increase the difficulty level of acceptable literacy. Thus, older citizens, who have every reason to believe they are adequately literate for all practical purposes, may discover that the threshold has moved away from them. Learners of English as a second/foreign

language are, likewise, caught in a constantly changing environment and may never be able to catch up with the advancing level. The point is that the relationship between some culturally defined standard of literacy and the non-linguistic forces (economic, social, political, etc.) operating on the culture is not well understood and needs further study (Bruthiaux 1992, Kaplan 1990).

### 3.5.4. Special Varieties

There are two distinct special varieties that have emerged in the recent past: on the one hand, a whole discipline defined as "English for special purposes" has emerged to teach defined narrow varieties necessary to accomplish specific but limited objectives (e.g., reading technical literature): on the other hand, the varieties requisite to special populations like the blind (e.g., languages like Braille) have been recognized.

#### 3.5.4.1 English (and other languages) for special purposes

Some of New Zealand's non-English speaking population would, conceivably, profit from a variety of courses in English for special purposes (e.g., English for auto mechanics, English for hotel personnel, English for health professionals) especially in conjunction with vocational training. If appropriate English-as-a-second-language instruction were provided for an adequate period of time, then it would be possible to top off such instruction with an additional period of special purposes instruction in a vocational training setting. (By the same token, it might be possible to mount special purposes instruction for other language learners; e.g., New Zealanders learning, say, Japanese might receive "Japanese-for-tourism" special purposes instruction following extended basic instruction in the language.) The point is that targeted special purposes instruction can turn marginal general purposes language proficiency into a marketable proficiency. Whether special purposes language instruction is exportable is a different question; it may be the case that the requisite general proficiency is inadequate in the typical FL situation to permit special purposes instruction to achieve its objectives, and the offering of special purposes instruction in off-shore contexts may build expectations which cannot be met solely by such instruction.

#### 3.5.4.2. Special populations

Unlike New Zealand Sign Language, which is a unique language employed by the deaf community, different from American Sign Language and Australian Sign Language, other disabled communities like the blind community can employ special varieties of English, like Braille. It is important that instruction in these varieties should be made widely available to these special populations throughout schooling and into adulthood to assure proficiency in a variety that gives access to literacy to a population otherwise cut off from standard script.

### 3.5.5. Translation and Interpretation

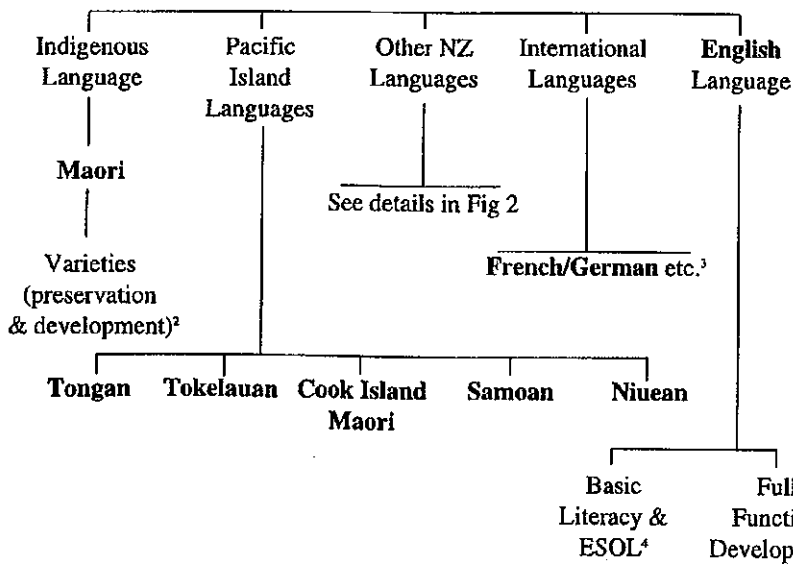
As New Zealand's technical needs in international trade increase, as the number of non-native speakers of English (especially in the lower socio-economic levels of society) increases, and as the needs of various populations in special registers increase, there is a growing need for high-level translating and interpreting services

based on English (because English is the *de facto lingua franca*). In the judicial system, for example, litigants who are not fluent in English – the language of the judicial system – should have available to them adequate translation services before, during, and after trials, By the same token, and perhaps even more critically, non-native speakers of English need access to appropriate translation services in the delivery of medical and other health-related services. As in many other essentially monolingual countries, there is a tendency to depend on "pick-up" translation in health services delivery situations, so, for example, a member of a hospital's janitor staff may find him/herself interpreting in a life and death situation solely by virtue of the fact that the individual happens to be a native speaker of the patient's language and not on the basis that the individual understands the medical complexity of the situation nor has a thorough grasp of the choices really available to the patient in the context of informed consent or of the question of confidentiality. But translating and interpreting services are needed not only in such critical areas as the judicial or the medical; the normal course of international business often requires the rapid translation of letters, the preparation of bilingual contracts, and not uncommonly simultaneous translation in negotiating environments. These needs can best be served by a readily accessible cadre of qualified interpreters and translators, properly trained and certified, appropriately compensated, and duly sensitive to questions of confidentiality (Douthett *et al.* 1990, Setefano 1991).

### 3.5.6. Summary

In sum, New Zealand's language situation may be crudely represented in Figures 1 and 2 below:

Figure 1: New Zealand Languages[1]



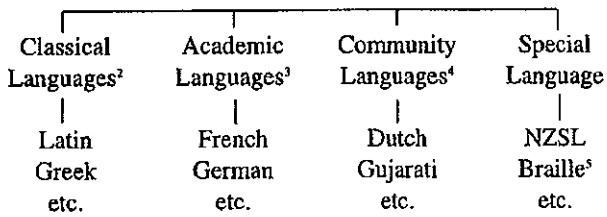

Notes to Figure 1

1. In each case developed in this small taxonomy, there are two basic issues: one issue relates to the ability to function in the language and subsumes such subsidiary issues as pronunciation, spelling, punctuation, syntax, morphology, etc., and the other basic issue implicates cultural understanding and subsumes pragmatics, semantics, literacy text

development, (i.e. the development of texts for the teaching of literacy to aliterate individuals) and literary comprehension. In the case of the academic and international languages, it should be understood that full functional ability is unlikely to develop exclusively in the classroom but will require the allocation of instructional time in a community in which the given language is actually spoken; in the case of Maori, Pacific Island languages, community languages (see Figure 2 below), and English, supporting communities in which those languages are used exist in New Zealand and such communities can function as essential reinforcement of classroom based instruction.

2. There is an important question, to be decided by Maori people, whether it would be desirable to develop a "standard" written and spoken Maori for wider use in New Zealand, for instructional purposes, and for official texts, such a standard form having no necessary or immediate impact on the preservation of spoken regional/tribal (Iwi) Maori varieties. Presently, standards are set by, for example, English-language newspapers.
3. This set of languages includes the generally recognized "international" languages: Chinese (Mandarin), English, French, German, and Russian. English as an international language is not the property of native speakers; it operates in a limited set of registers (never the personal ones) and it serves a limited set of functions (generally public ones).
4. The issue is not exclusively an ESOL problem: on the contrary, some speakers of English as an alternative language, both New Zealand born and immigrant, may have achieved high levels of English literacy while, at the same time, some New Zealand born speakers of English as a first language may lack even basic literacy. It must be recognized that, while methodologies for first-vs. second-language instruction may vary, the underlying problems are characteristic of the entire population without reference to place of birth, visa status, self-identified first language, or educational level. At the same time, it must be recognized that language proficiency is only one of a syndrome of issues which affect educational achievement – issues like poverty, diet, health, home environment, parental attitudes toward education, etc.

Figure 2: Other New Zealand Languages[1]




Notes to Figure 2:

1. Originally, in earlier versions of this research, this category had been labelled "Foreign" languages. It was pointed out to the author that the term *foreign* might be resented by New Zealanders who speak various community languages, and that designating these languages as "foreign" (when Pacific Island languages are not so designated) might imply that they were less important. On the contrary, it must be noted that some language communities, like Cantonese or Gujarati for example, are much larger than some of the Pacific Island language communities. I confess to my inability to devise a label more accurately descriptive than "Other New Zealand languages."
2. This set of languages may also include other classical languages such as Anglo-Saxon, Classical Hebrew, Old Church Slavonic, Old High German, Old Icelandic, Sanskrit, etc., as well as languages which serve a basically religious function (e.g., Church Greek, Classical (Koranic) Arabic). It is important to understand that the treatment of these languages in a planning environment will vary from the treatment of living languages.

3. This set of languages may include less-frequently taught languages which may be primarily important for economic purposes (e.g., Modern Arabic, Bahasa Indonesia, Bahasa Malaysia, Korean). The academic French and German taught in New Zealand is not the same as the migrant varieties of the languages or the metropolitan varieties.
4. This set of languages may include other "community languages" like Cantonese, Cambodian, Hmong, Italian, Modern Greek, Vietnamese etc. It is important to note that languages contained in this and the academic languages set may in fact overlap, but will be differently treated depending on which of the two categories they are deemed to fall into.
5. It may be argued that Braille is not a language but rather is merely a variant notational system for transcribing English, French, etc. There is increasing evidence that Braille is not merely a notational system but, whether it is or not, it must be given adequate consideration across a range of related sectors (e.g., publishing, education, international travel).

### *3.6. Review of the Language Situation*

This brief discussion suggests that there are a large number of languages and varieties in use in New Zealand – perhaps as many as thirty-five or forty languages, including several varieties of Chinese (Cantonese/Mandarin), English (International/British/New Zealand), French, and German. The discussion, however, is not based on hard evidence, but largely on observation and on a number of more or less notional discussions of various New Zealand languages with individuals representative of various ethnic communities. There is a critical need to collect real empirical information on the language situation. At present, decisions at all levels of the governmental system (including the public education sector) must be based on imperfect information (sometimes on no information at all), and decisions tend to be organized from the top down. What New Zealand needs is a large scale sociolinguistic survey to determine who speaks what to whom under what circumstances to achieve what ends. And the need for a sociolinguistic survey implicates the need for an organization not only capable of conducting such a survey but also of interpreting the results and of maintaining the currency of the data base. No such organization presently exists (or is presently contemplated) in New Zealand.

Whatever the actual number of languages involved, the language situation is a complex one. That complexity is increased by the *de jure* status of Maori and the *de facto* status of English. It is clear that language planning needs to take account of the real linguistic heterogeneity of the nation, to determine what the nation can afford (and what it can afford to sacrifice) in terms of the maintenance of its linguistic diversity, and to develop a set of rational priorities for maintaining and extending languages that are, for whatever reason, important to New Zealand's future as well as to its past.

## 4. The Economic and Political Environment

In the mid-1980s, New Zealand found itself in a situation in which its foreign debt was so large and its annual debt service so great that it was unable to borrow further in the international monetary community. It had to get its financial house in order. (In all fairness, the economic problem was equally attributable to other causes – e.g., the loss of export markets after Britain's entry into the EC. Much of the borrowing that took place was required to close the gap in government revenues and to continue to finance the generous levels of welfare services for which New

Zealand was internationally recognized and to which New Zealanders had become accustomed. Indeed, it was the first Lange government, under the leadership of Finance Minister Roger Douglas, which installed "Rogernomics", cutting social welfare entitlements, which resulted in the unpopularity of the Labour government and caused the Conservative swing among the electorate.) Given the state of its economy, it was not surprising that the government became formally Conservative in the late 1980s. The election of the first Lange government occurred not very long after the election of the Thatcher government in Britain and the Reagan administration in the United States. Conservative governments in the Anglo-European states have tended to favour low taxes, limited (small, compact) government, and limited governmental regulation/intervention. New Zealand has been no exception to this general conservative trend, which had occurred to varying degrees under the Thatcher government in Britain (involving a good deal of privatization and some tax cutting) and the Reagan administration in the United States (largely limited to tax cutting). (Though both Thatcher and Reagan talked about reducing the role of government, neither actually did anything significant in that context.) In fact, New Zealand has gone further along these lines than either of its alleged models.

In the late 1980s, New Zealand disassembled its long-standing Department of Education (and other national departments as well) in favour of a ministerial structure on the theory that cabinet officers ought to have greater power, ought not to be constrained by the intransigence of entrenched civil service officers, and ought to be able to implement their individual visions for the sector with which they were charged (Ball 1992, Boston 1991). The ministerial structure was seen as leaner (and indeed staffing has been significantly reduced in a number of sectors), as more egalitarian (that is, it has been assumed that "policy analysts" (middle-level bureaucrats) are interchangeable, so that personnel can be shifted across government agencies (ministries) as needs alter), and as more productive (on the grounds that a Minister would "buy" only those functions s/he wished to have performed, i.e., those perceived by the Minister as having high priority). In general, this was a cost cutting measure, both to reduce direct expenditure by government and at the same time to reduce indirect expenditure by reducing government intervention/regulation, but it also constituted a major shift in direction, because New Zealand had previously been universally perceived as a well-developed socialist state. In the education sector, this policy has led to devolution of schools and to the application of the concept "user pays" (also prevalent in other "social-welfare" sectors of society); thus, parents pay directly for services they want for their children and they pay their fees not to government in the form of taxes, but to independent schools which are earning their own way in a competitive market. Schools compete; better schools attract more students, better teachers, and more funds, and become excellent while weaker schools go out of business. In sum, the notion is deregulatory, removing governmental control, and subsidy from education (and other social services) and causing providers to compete in the open market[6].

In 1992, when the author was working in New Zealand, the government was still struggling with what precisely to devolve and what to keep. The Ministry of Education was, for example, involved with a range of more-or-less competing activities touching on language education:

- it had published requests for proposals to develop syllabi in Samoan and a few other languages;
- it was investing in the creation of a new English syllabus, presumably responsive to current needs;
- it was engaged in a major activity to develop a new national curriculum (Ministry of Education 1991), and
- it had permitted the development of a "green paper" for a national languages policy (Waite 1992).

While all of these activities were centred in the Ministry of Education, they were, to all intents and purposes, separate and distinct efforts with little or no cross-talk among them. And it remained unclear at the time (mid-1992) which of these activities government ought to be undertaking, this discussion being somewhat clouded by an underlying attempt to differentiate between policy and philosophy on the one hand and implementation and operation on the other. But it was already fairly clear the languages policy development was not a high priority in the Ministry of Education and was certainly not a priority at all in other government agencies.

It can be claimed that the redirection of the education sector was beset with certain contradictions:

- On the one hand, education is such an expensive commodity that it demands continuity across time; it is not productive to change direction at short time intervals because adequate evaluation requires long-term implementation. On the other hand, Ministers of Education have the right to implement their vision, regardless of the chronological (or logical) consistency of that vision in relation to whatever happens to be in progress.
- On the one hand, the most important social obligation of any community is the education of its youth; education of the young constitutes *the* major human capital investment of any society in its future. On the other hand, the notion "user pays" places responsibility for the education of the next generation squarely and exclusively in the hands of the parents; it forces parents to make individual, often uninformed, decisions that will have the most profound impact on the future manpower pool of the community.
- On the one hand, schools become empowered to buy on the open market whatever they need to provide high quality education. On the other hand, schools inevitably are forced to draw on a local supplier-sector and thus to lose the savings possible through economies of scale.

There are many such contradictions in the current system. Clearly, language education is likely to suffer in such an environment for several reasons. First, the pool of available language teachers is always smaller than the language needs of society, and the pool is uneven across languages (Gomez de Matos 1992). Both teachers and students have rights which must be protected if language is likely to be delivered and received in a quality environment; because language is a universal phenomenon, everybody is an "expert", and the seriousness of language teaching and learning is misperceived and underestimated (Fédération Internationale des Professeurs de Langues Vivantes 1992). Thus, "popular" languages have access to larger sectors of the market, while "unpopular" languages – even though they

may be socially or historically important – have access to smaller segments of the market. Because languages like Japanese are likely to be perceived by parents as having great prestige at the moment and as holding out the promise of employment for their offspring, while languages like Samoan are perceived to be the primary concern of the Samoan community and not of the total population, Japanese is likely to attract a far larger share of the market than Samoan, while the teacher-pool for both languages is likely to remain well below the needs-level (though for different reasons) and is unlikely to be brought to minimum satisfactory levels (because teachers, like students, seek their individual long-term economic good). It can be said that the government has abrogated its responsibility to a minority community (the Samoan community in this illustration, but probably all of the Polynesian language speaking communities and a number of other stigmatized communities as well), but it can also be said that since it is not the case that New Zealanders across the spectrum of the society want to learn Samoan (or other non-prestige languages), the maintenance of Samoan (or other "community" languages) is indeed the responsibility of the respective minority communities and should not be supported with national government (tax-based) funding – as a playing out of the principle "user pays." But the Samoan community has (and most of the other minority communities have), for a variety of complex reasons, less internal resources at its (their) disposal than does the community which is interested in having its offspring learn Japanese. The result is that, as the old depression era song says, "...the rich get richer and the poor get children." It can be claimed that this is a perfect working out of the capitalist paradigm; at the same time, it can be shown that government is failing to meet the social requirements of the most needy sectors of the community. These are not issues that an alien can usefully address; they are emotionally charged, and they implicate the most fundamental priority choices of the society.

## 5. Language Planning

When, in the very late 1970s, IndoChinese refugees began to arrive in New Zealand in significant numbers, an interest in language planning began to emerge. A number of committees, study groups, colloquia, and symposia were convened through the late 1970s and early 1980s, and a number of reports of various sorts were prepared (see, e.g., Kaplan 1980). Many (but not all) of the reports urged the development of a national languages policy (National Language Policy Secretariat 1989). When, in the mid and late 1980s, Australia developed a national languages policy (Lo Bianco 1987, Australian Advisory Council on Languages and Multicultural Education 1991), interest in a languages policy in New Zealand increased (a playing out of a long-standing love/hate relationship between the two polities). Finally in the late 1980s the then Labour government agreed to take language policy aboard as one of its concerns, and the current Conservative government did not abandon that commitment when it took office at the end of the decade. (See Peddie 1992 for a more detailed summary of events.)

In 1990, the Ministry of Education commissioned two activities; on the one hand, it let a contract to Dr. Roger Peddie of the University of Auckland to study

developments in the state of Victoria, Australia, as they compared with developments in New Zealand (Peddie 1991), and on the other hand, the Ministry of Education employed Dr. Jeffrey Waite to develop a national languages policy paper for the Ministry (Waite 1992). It was decided to invite a specialist in language policy from the United States under the terms of the Hays-Fulbright program; the assumption was that the Peddie and Waite reports would be delivered to the Minister of Education for his review before the 1991 Christmas holidays, that the Minister would review and release for dissemination at least the Waite report, and that the foreign expert would then be able to work with the wide commentary anticipated as the result of the broad dissemination of one or both of those reports when s/he arrived in early 1992. Both researchers provided their reports on schedule (see Peddie 1991, Waite 1992), and the foreign expert was selected and scheduled to arrive in mid-March 1992. Unfortunately, for unknown reasons, the Minister of Education did not complete his review by early January 1992, and when he did eventually complete his review he found himself uncomfortable with at least the Waite document. (It is not clear whether he ever read the Peddie interim report.) Thus, when the foreign expert (the author of this paper) arrived in mid-March, it was unclear how he would spend his five month contract time. In an early meeting with the Minister, within days of the arrival of the foreign expert, it was decided to appoint a "Ministerial Advisory Committee" to review the Waite report and advise the Minister with reference to its disposition. The Advisory Committee (three representatives of various sectors of New Zealand society – industry, the schools and the Maori community – with the foreign expert serving as convener) did its work very quickly and within approximately three weeks forwarded a report to the Minister urging him to release the Waite report for wide dissemination. The Advisory Committee also recommended the removal of the language planning activity from the Ministry of Education on the grounds that a national languages policy ought to include a much larger and more representative segment of the government and the community, the establishment of some sort of national organization (four or five alternative models were suggested) to advise on language policy issues and engage in essential data collection on the grounds that available information was inadequate to the formulation of any policy or indeed to any decision-making. The Minister never formally responded to the recommendations of the Advisory Committee, but he did subsequently arrange to have the Waite report reviewed for political sensitivity (correctness?). At long last, in May 1992, he agreed to have the Waite report released and widely circulated for comment. It took some weeks to produce the report in several thousand copies and to deal with various logistic issues relating to the dissemination of the report. Finally, in the last week of June 1992, the report was in fact widely disseminated (but largely to an education-sector mailing list), with a cover letter from the Minister inviting comment by 1 October. The foreign expert's contract terminated on 1 July, and he left the country, having filed prior to his departure a final report with the Ministry of Education covering his activities from mid-March to the end of June (Kaplan 1992).

The current state of the national languages policy effort is not known, though it is known that a meeting of several language teacher associations was held in Auckland at the end of August and that there was a scheduled session set aside for the discussion of the Waite report and of the national languages policy issue more

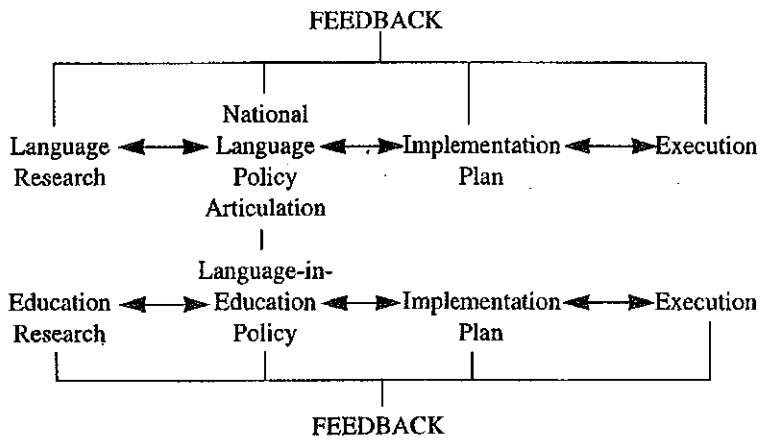
broadly. The Ministry of Education has planned to convene (in the fiscal year beginning 1 July) an interagency committee to deal with the various issues raised by the report, and Dr. Waite's appointment was continued in the Ministry to coordinate the effort. It is known that a mechanism was developed within the Ministry to deal with responses to the Waite report received by the Secretary of Education (the Ministry CEO) before the 1 October deadline.

## 6. What might have been

That there is a probable need for languages planning in New Zealand is to some degree demonstrated by the complexity of the language situation and by the paucity of data regarding the uses of various languages in New Zealand[7]. That the directions of such a policy remain obscured is demonstrated by the number of uncoordinated activities on-going within the Ministry and in various ethnic communities. That there is a broad interest outside government, particularly in the language-teaching community, is demonstrated by the number of people who have continued, over the twenty-odd years since the idea was first bruited about, to come together in various configurations in the language-teacher organizations, under the auspices of the Institute of Policy Studies at Victoria University, under the auspices of the New Zealand Council for Educational Research, and in other less structured and unstructured contexts. (It is important to remember that New Zealand is a relatively small country and that, at least in the academic sector, people tend to know each other well, to have been to school together, and to meet socially on an occasional basis.)

In a sense, while any movement towards a national languages policy is welcome, it is unfortunate that what limited current movement there has been has occurred in the Ministry of Education. It is demonstrable that language-in-education planning should follow from broader national planning, not lead it, as is suggested in Figure 3:

Figure 3: Language Policy Development Model




The focussing of the language policy effort in the Ministry of Education has meant that great segments of New Zealand society have not really been consulted on the complex questions implicated by a national languages policy. This is not to

condemn the Waite report (1992 qv) but rather to point out that a single individual working out of the Ministry of Education under rather tight time constraints and without adequate staffing, funding, or outreach is unlikely to be able to conduct a national sociolinguistic survey or simply to be able to contact many sectors outside the education purview. Language planning is hardly a new discipline, and a great deal of guidance is readily available in the literature (Annamalai, Jernudd and Rubin 1986; Baldauf and Luke 1990; Haji Omar and Noor 1981; C Kennedy 1983; Singh and Srivastava 1987).

In its report to the Minister, the Ministerial Advisory Committee suggested a number of sectors (13 to be precise) which ought to be consulted in the development of a national policy:

| CONSUMER SECTORS ◄——— | | ———► PROVIDER SECTORS |
|---|---|---|
| Commercial sector | Local Government sector | Communication sector |
| Foreign Aid sector | Military sector | Education sector: |
| Foreign service sector | Police/prisons sector | i) Primary/secondary education |
| General governmental sector | Tourism sector | ii) Tertiary education |
| Health sector | Trade union sector | iii) Adult education |
| Judicial sector | | iv) Vocational education |

Although the Advisory Committee did not explicitly specify contact with the various ethnic communities, in the thinking of the Advisory Committee members such exploration is so obvious it need not be specified. The Advisory Committee was concerned that decisions were likely to be made without adequate consultation outside the educational/governmental sectors, that decisions were likely to be top-down simply because all the planning was being done at the governmental level, and that the Ministry of Education lacked the resources and the outreach to influence policy development in any sector of society beyond government, or perhaps even beyond the education sector. Although the notion of creating an interagency committee emerged after the Advisory Committee had completed its duties, it is likely that the Advisory Committee would have opposed such a solution on the same grounds that it opposed the exclusive development of languages policy in the education sector.

The Advisory Committee did, on the other hand, recommend the creation of a free-standing body charged with the research necessary to develop a data base, with the technical analysis of linguistic and social information, and with an advisory role to government. It is recognized that only government can make policy decisions, but that government probably lacks the information – and expert knowledge to use such information – or the time to develop the information and expert knowledge. A body of the kind recommended by the Advisory Committee should operate at arm's length from government, in order to be able to continue its work without reference to the political needs of any given administration. The New Zealand National Languages and Literacy Institute (NZNLLI – as it might be designated) should be charged with the regular collection and interpretation of linguistic and sociolinguistic information, with the provision of advice to various ministries on language issues, with the development of criteria for interpreting and translating services, for language assessment in the education sector and elsewhere, with the articulation of an advertising campaign to

raise language awareness in the general population, and with the development of current information on New Zealand's requirements in the context of international trade. In sum, the NZNLLI should serve as a catalyst in the broad area of language awareness. It should work cooperatively not only with various ministries of government, including those specifically charged with Maori and Pacific Islander affairs, but with international corporations, with trade unions, with consumers of language training, and with providers of all aspects of language training throughout the society.

With respect to the endangered languages (the Polynesian languages, including indigenous Maori), it is unclear what might be done to arrest the decay of those languages (N Benton, 1989; R Benton, 1991a; R Benton, 1986). It is well established in the theoretical framework underlying language planning that languages die when:

- parents, for whatever reason, do not engage in intergenerational transmission of the language in question;
- the language loses registers to another language, thereby becoming increasingly constricted in the real-world functions served by the language in question; and
- younger individuals are drawn away from the community by economic pressures.

All three of these criteria are met in New Zealand with respect to Maori and the other Polynesian languages. Some parents see no future in the languages and do not bother to transmit them to their children, believing that English has greater promise for the future of the children. Maori and the other Polynesian languages serve an increasingly constrained set of registers; that is, most things in New Zealand get done in English, and Maori and the Polynesian languages are being gradually constricted only to certain limited ceremonial functions. There is a curious irony in the Maori and Pacific islander situation in the sense that many Maori and Pacific Island people have been converted to Christianity, but Christianity is not a social structure through which Maori and more general Polynesian values can be retained; while Maori is used in Christian religious worship by Maori people, Christianity represents a set of spiritual and social values at odds with *Maoritanga*. Finally, the economic structure of New Zealand is such that young Maori are pulled away from ethnic communities in order to find employment, that having left the community they tend to marry outside the community, and that they subsequently tend to raise their offspring outside the community structure.

It is not easy to determine whether these tendencies can be reversed. Even if they can be reversed, the outcome is likely to produce not a community of Maori speakers but a community of English speakers who can perform a limited number of functions in Maori. But the fact remains that, in the current environment, not enough is known to understand how the problem should be treated. The longer information remains sparse, the more difficult it will be to determine an appropriate strategy to treat the problem. It is, again, well established that there is a "critical" point – a point of diminishing returns – beyond which the revival of a language becomes highly improbable; it is not known whether Maori and the other Polynesian languages have reached that critical point.

## 7. Summary and Conclusion

What then has happened to the Conquest of Paradise? For Maori people, Paradise existed in retrospect in a time before English settlers came to New Zealand. That condition is certainly no longer re-attainable, at least in part because, like all mythic conditions, it never really existed, a situation that does not in any sense inhibit the aspiration. Like it or not, the history resulting from the arrival of English settlers in New Zealand is irreversible. New Zealand remains a predominantly English-speaking polity. Although the rights of Maori people are protected by law, the intermingling of the two populations makes identification as a Maori person increasingly difficult, and makes the areas over which the protected rights exist more tenuous. And the Maori language is marked by the three conditions that may signal language death:

- Maori parents do not invariably pass their language intergenerationally; indeed, the middle generation is extremely attenuated, so that in many cases only grandparents are capable of transmitting the language;
- Maori language has lost a very large number of registers to English, and although religious matters may be carried in Maori, it is Christian, not Maori, values that are so transmitted;
- Basic economic structures are in the hands of English speakers; it is not necessary to speak Maori in order to get and hold a job, and young people are regularly drawn away from Maori communities by the pull of economic and accompanying linguistic forces while family ties may not be strong enough to counter this force.

The rights of speakers of other Polynesian languages are not so carefully protected by law – only those rights which accrue to all citizens are protected. While *Te Reo Maori* is recognized as a treasure of the Maori people under the provisions of the Treaty of Waitangi, the languages of other Polynesian groups are not so recognized, nor are the languages of other non-English speaking minorities. It is unclear whether the languages of Indo-Chinese residents, the newest arrivals, will survive beyond the second immigrant generation (Holmes *et al* 1992). Even if those languages do survive, it is unclear what functions they will fulfil in New Zealand and in what registers they may operate effectively.

English, on the other hand, is not in any real danger, but the arena in which English functions is marked by confusion over the prestige variety, by increasing literacy issues, and by a failure to provide necessary services even in English to language minority communities.

If Paradise is English-speaking, then that paradisical condition is very near being achieved, but if Paradise is a multilingual, multicultural society, there is far to go; indeed, it is entirely possible that the goal cannot be achieved. New Zealand will, of course, survive; that isn't the question. Rather, will New Zealand survive at the cost of its diversity? Only New Zealanders can answer the question, but they must answer it soon or the answer may become irrelevant because the disease is too far advanced. Language death – like any other death – is not a pretty thing to watch.

> ...Farewell happy fields
> Where joy forever dwells: Hail horrors, hail
> Infernal world, and thou profoundest Hell

Receive thy new possessor: one who brings
A mind not to be changed by place or time.
The mind is its own place, and in itself
Can make a Heav'n of Hell, a Hell of Heav'n.
(John Milton, *Paradise Lost* Bk.i , 1.247.)

## Notes

1. A shorter version of this paper was delivered at the annual 1993 TESOL conference in Atlanta, GA.
2. Language education is, to some extent, still captured by an older paradigm; when language education was first introduced into the academy – in the medieval university – the languages at issue were the classical languages (essentially Classical Greek, Classic Hebrew, Latin, and Sanskrit). These were dead languages, with fixed syntax and lexicon, and a limited, frozen inventory of texts. The objective of language study was not communicative proficiency (since there was no real community with which to communicate), but rather was access to the thought, culture, and art of a dead civilisation. Under those circumstances, it was reasonable to select the most intellectually able students for instruction and to employ a grammar-translation methodology. When modern languages were introduced into the school curriculum in the late 19th century, the original instructional paradigm was retained; language instruction was reserved to the most intellectually able students, the grammar-translation methodology was employed and the object of instruction was access to the canonical literature of the language. Thus, communicative competence, because it was not an objective, was seldom achieved. (There is some evidence that, if children learn anything at all in school, they learn only what they are taught.) Though language curricula have gradually become more concerned with communicative competence, the older literary bias remains, and the students coming out at the end of a period of instruction may still have largely a literary orientation. In programs directed at business functions, this literary bias is misplaced, because students emerge with a set of skills which do not contribute to their ability to function in the domains for which they ostensibly have been trained.
3. Through most of human history, the development of science and technology has proceeded at a slow and stately pace. An ordinary individual could normally live a full and happy life without ever encountering any sort of threatening scientific and/or technological change. For the most part, scientists were, in the past, amateurs and dilettantes, working at science out of curiosity, and supporting that work either through personal wealth or through the generosity of wealthy patrons. It was not until well into the industrial revolution that industrialists and businessmen recognized the need to harness science to technology for profit, and it was not until quite recent times that the class of professional scientists emerged, working at science on a daily basis, largely in academic institutions but also in industry itself. Even more recent has been the emergence of the great "research university" and the corporate "think tank" which pursue particular directed (funded) scientific research and which have been driven largely by direct funding either from industry or from government. At present, a number of major industries (the aerospace industry, the automotive industry, the health industry, the so-called "knowledge industry", the pharmaceutical industry, etc, – industries in which the turn-around time from scientific breakthrough to marketable technological innovation is very short) maintain large "R&D" (research and development) sections of professional scientists. These R&D sections are in turn supported by a new professional class – information scientists and information managers – who funnel pertinent information to working scientists. The development of a class of professional scientists, of research venues for the pursuit of

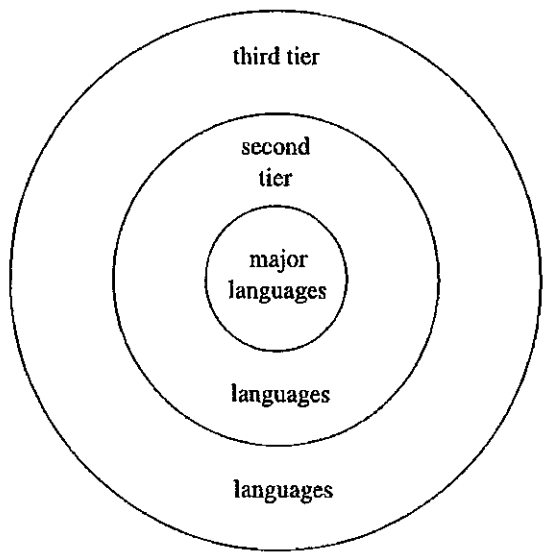
targeted science, and more recently of professional information scientists and information managers have had the most profound cultural and linguistic implications for polities in which this pattern has developed.

4. It is important to note that other cartels – OPEC for example – deal with a diminishing commodity (i.e., petrochemical resources diminish with use and new sources constantly have to be sought at great expense), while the English-based information cartel deals with a constantly expanding commodity because information expands as it is used.
5. At the same time, it is important to recognize that the emergence of these "major" languages constitutes a threat to the existence of "smaller" languages; as more resources are put at the disposal of the major languages, fewer resources remain available to the smaller languages – and smaller is a relative construct; in this context, it appears to mean any language which is not important to the transmission of information at more than the local level, so that Breton and Basque are more directly threatened than Dutch or Danish, but the conditions of English and French are categorically different. One may conceive of the situation as consisting of a set of concentric circles in which the innermost circle encounters no threat, the middle circle some threat, and the outer circle the greatest degree of threat. Individual languages may move across circle boundaries. In a European context:

   Major languages = English, French, German, etc.
   Second tier languages = Danish, Dutch, Hungarian, Italian, etc.
   Third tier languages = Basque, Breton, Frisian, Samish, etc.

   In a New Zealand context:

   Major language = English
   Second tier language = Cantonese, Dutch, Modern Greek, Italian, etc.
   Third tier languages = Cambodian, Hmong, Maori, Nuean, Samoan, Tokelauan, Tongan, etc.




6. This is something of an oversimplification of monetary philosophy; I believe my summary to be accurate, but my concern lies less with philosophy than with its outcome.
7. Obviously the fact that something is complicated does not justify planning – witness the economy; a paucity of data requires the collection of data providing a planning effort has been agreed to. Language planning requires data; the absence of data does not require language planning. Given this author's bias, it is the case that language planning is necessary at least to save the endangered languages, but perhaps also to attend to serious

social problems arising from the absence of coherent language services in English. But whether language planning should be a high priority in a constrained economic environment is a decision that can only be reached by the populations whose monies will be spent to do the planning – and, incidentally, who benefit from the outcomes of planning. Because the author is not a New Zealand citizen, is not a New Zealand tax payer, has not educated his children in New Zealand, and is not Maori, he can advance a more or less logical defence of the importance of language planning; he cannot do more. The decision rests, ultimately, with New Zealand voters.

## Bibliography

AGB/McNair. 1992. Survey of demand for bilingual and immersion education in Maori: A report to the Ministry of Education. Wellington: AGB/McNair.

Annamalai, E., B.H. Jernudd and J.Rubin. eds 1986. *Language planning: Proceedings of an institute*. Mysore: Central Institute of Indian Languages.

Australian Advisory Council on Languages and Multicultural Education. 1991. Language planning in Australia: A report. *New language planning newsletter*. 5.3.3

Baldauf, R.B., Jr and B.H.Jernudd. 1983. Language of publication as a variable in Scientific communication. *Australian review of applied linguistics*. 6.1.97-108.

Baldauf, R.B., Jr and A. Luke. eds. 1990. *Language planning and education in Australasia and the South Pacific*. Clevedon, Avon: Multilingual Matters.

Ball, C. 1992. Ladders and links. Unpublished paper presented 21 January at the International Conference on "Qualifications for the 21st Century," Victoria University of Wellington.

Bates, D.L. 1991. Maori language: Some observations upon its use in criminal proceedings. *New Zealand law journal*. February: 55-60.

Bell, A. and J. Holmes. eds. 1990. *New Zealand ways of speaking English*. Clevedon, Avon: Multilingual Matters.

Benton, N. 1989. Education, language decline and language revitalization: The case of Maori in New Zealand. *Language and Education* 3. 2. 65-82

Benton, R.A. 1981. *The Flight of the Amokura*, Wellington: New Zealand Council for Educational Research.

Benton, R.A. 1986. Schools as agents for language revival in Ireland and New Zealand. In B. Spolsky, ed. *Language and Education in Multilingual Settings*, Clevedon, Avon: Multilingual Matters

Benton, R.A. 1991a. Notes on the case for Maori language television. *New Language Planning Newsletter*. 5.4. 1-4.

Benton, R.A. 1991b. "Tomorrow's Schools" and the revitalization of Maori: Stimulus or tranquilizer? In O. Garcia, ed. *Bilingual Education: Focusschrift in honour of Joshua A Fishman on the occasion of his 65th birthday*. Vol. 1. Amsterdam: John Bejamins. pp 136-147

Bolger, Rt. Hon. J.B (Prime Minister). 1992. Address at the Asia 2000 Conference, Wellington Parkroyal Hotel, 28 May. Xerox.

Boston, J. 1991. The theoretical underpinnings of public sector restructuring in New Zealand. In J. Boston, J.Martin, J.Pallot, and P.Walsh, eds. *Reshaping the state: New Zealand's bureaucratic revolution*. Auckland: Oxford University Press.

Brain, J., P.Dugdale, and S. Logan. 1992. New Zealand sign language NZSL, Australian sign language ASE and total communication: What are they? *Communicate: The official journal of the National Foundation for the Deaf*. 2.5:33,35.

Bruthiaux, P. 1992. Literacy and development: Evidence from the Pacific rim and beyond. Ms. Los Angeles: University of Southern California. (Course paper, Linguistics 677, Fall 1991 semester.)

Coulmas, E. ed. 1992. *The economics of language in the Asian Pacific.* (Special issue, *Journal of Asian Pacific Communication*, 2.)

Douthett, M., et al. 1990. *Report of the working party on interpreting services.* Auckland: Auckland Area Health Board. (Cartwright implementation Taskforce.)

Dunn, A., N. Pole, and J. Rouse. 1992. *The education sector workforce.* Wellington: Ministry of Education.

Fédération Internationale des Professeurs de Langues Vivantes 1992. Human language rights. *FIPLV world news.* 58.1-2.

Fishman, J.A. 1991. Maori: The native language of New Zealand. In J.A Fishman. *Reversing language shift.* Clevedon, Avon: Multilingual Matters. 230-251.

Gomes de Matos, F.1992 Foreign language teachers' rights: A plea for world documentation. *FIPLV world news* 58.3.

Government of New Zealand. 1987. *The Maori Language Act.* Wellington: Government Printer.

Grace, R. 1991. *Promoting Maori language and culture within current structures and funding.* Wellington: Ministry of Education.

Haji Omar, A. and N.E.M. Noor. eds. 1981. *National language as medium of instruction.* Kuala Lumpur: Dewan Bahasa dan Pustaka, Kementerian Pelajaran Malaysia.

Hawley, C. 1988. Educational services to immigrants and refugees 1988. Wellington: Department of Education. (Appendix IV to an unknown document.)

Hirsh, W. ed. 1987. *Living languages: Bilingualism and community languages in New Zealand.* Auckland: Heinemann.

Hirsh, W. 1990. *A report on issues and factors relating to Maori achievement on the education system.* Auckland: Ministry of Education.

Holmes, J., M. Roberts, M. Verivaki, and 'A. 'Aipolo. 1992. Language maintenance and shift in three New Zealand speech communities. *Applied Linguistics* 14. 1. 1-22.

Iosia, P.S. 1992. How do you say...? Wellington: Pacific Island Management Development Course, Placement project, Ministry of Education. (Unpublished report)

Kaplan, R.B. 1980. *The Language Needs of Migrant Workers.* Wellington: New Zealand Council for Educational Research

Kaplan, R.B. 1981. The language situation in New Zealand. *Linguistics Reporter.* 23. 9. 1-3

Kaplan, R.B. 1990. Literacy and language planning. *Lenguas Modernas.* 17. 81-91.

Kaplan, R.B. 1992. *New Zealand National Languages Policy: Making the Patient More Comfortable.* (A Report to the Policy Division of the New Zealand Ministry of Education). Wellington Ministry of Education.

Kaplan, R.B. and P.Medgyes. 1992. Discourse in a Foreign Language: The example of Hungarian scholars. *International Journal of the Sociology of Language.* 98. 67-100.

Karetu, T.S. 1991. Te Ngahurutanga: A decade of protest, 1980-1990. In G.McGregor and M.Williams, eds. *Dirty silence: Aspects of language and literature in New Zealand.* Auckland: Oxford University Press.

Kennedy, C. ed. 1983. *Language planning and language education.* London: George Allen & Unwin.

Kennedy, G.D. 1982. Language teaching in New Zealand. In R.B. Kaplan, et al. eds. *Annual review of applied linguistics,* 2. Rowley,Ma: Newbury House 189-202.

Kennedy, G.D. 1988. The learning of English in New Zealand by speakers of other languages. Unpublished paper presented to the First National Conference on Community Languages and English as a Second Language. Wellington.

Levett, A. and A. Adams. 1987. *Catching up with our Future: The Demand for Japan Skills in New Zealand.* Wellington: New Zealand Japan Foundation.

Lo Bianco, J. 1987. *National policy on languages.* Canberra: Commonwealth Department of Education.

McGregor, G and M. Williams,eds. 1991. *Dirty silence: Aspects of languages and literature in New Zealand.* Auckland, New Zealand: Oxford University Press.

Ministry of Education. 1989a. *Bilingual education in New Zealand.* Wellington: Maori and Island Division.

Ministry of Education. 1989b. *Maori education statistics, 1989.* Wellington: Department of Research and Statistics.

Ministry of Education. 1991. National Curriculum of New Zealand: A discussion Document (The). Xerox.

National Language Policy Secretariat. 1989. Toward a national languages policy for New Zealand. Wellington: Department of Education.

Panitchpakdi, S. 1992. Asia and New Zealand: Future prospects and linkages. Unpublished keynote address delivered at the Asia 2000 Conference, Wellington: Parkroyal Hotel, 28 May. Xerox.

Peddie, R.A. 1991. *One, two, or many? The development and implementation of languages policy in New Zealand.* Auckland: University of Auckland.

Peddie, R.A. 1992. Language and languages policy in New Zealand: Defining the issues. *English in Aotearoa.* 18. 40-50.

Setefano, S. 1991. *Wellington community interpreting service project.* Porirua: Housing Corporation of New Zealand.

Sexton, S. 1990. *New Zealand schools: An evaluation of recent reforms and future directions.* Wellington: New Zealand Business Roundtable.

Singh, U.N. and R.N. Srivastava. 1987. *Perspectives in language planning.* Calcutta: Mithila Darshan.

Smith, Dr. the Hon L.(Minister of Education). 1992. Asia 2000: Realizing the opportunities. Address delivered at the Asia 2000 Conference, Wellington Parkroyal Hotel, 28 May. Xerox.

Spolsky, B. 1987. *Report of Maori-English bilingual education.* Wellington, New Zealand: Department of Education.

Thomson, C. 1991. Use of Maori language factor funds in selected schools. *Research and statistics division bulletin.* (New Zealand Ministry of Education.) 3: 81-87.

Visser, H. and N. Bennie. 1991. The role and commitment of resource teachers of Maori. *Research and statistics division bulletin.* (New Zealand Ministry of Education.) 3: 36-54.

Wagemaker, H. (mss) Preliminary findings of the IEA reading literacy study: New Zealand achievement in national and international context. Typescript. 1992. Research and Statistics Section, New Zealand Ministry of Education.

Waite, J. 1992. *Aotearoa: Speaking for Ourselves: Issues for the development of a New Zealand language policy.* Wellington: New Zealand Ministry of Education.