Text Generation and Systemic-Functional Linguistics

Experiences from English and Japanese

Christian M.I.M. Matthiessen
and
John A. Bateman

Pinter Publishers, London

COMMUNICATION IN ARTIFICIAL INTELLIGENCE SERIES

Artificial Intelligence (AI) is a central aspect of Fifth Generation computing, and it is now increasingly recognized that a particularly important element of AI is communication. This series addresses current issues, emphasizing generation as well as comprehension in AI communication. It covers communication of three types: at the human–computer interface; in computer–computer communication that simulates human interaction; and in the use of computers for machine translation to assist human–human communication. The series also gives a place to research that extends beyond language to consider other systems of communication that humans employ such as pointing, and even in due course, facial expression, body posture, etc.

From Syntax to Semantics: Insights from Machine Translation, eds: Erich Steiner, Paul Schmidt and Cornelia Zelinsky-Wibbelt

Advances in Natural Language Generation: An Interdisciplinary Perspective, 2 volumes, eds: Michael Zock and Gérard Sabah

Text Generation and Systemic-Functional Linguistics: Experiences from English and Japanese, Christian M.I.M. Matthiessen and John A. Bateman

Further titles are in preparation

Text Generation and Systemic-Functional Linguistics

Experiences from English and Japanese

Christian M.I.M. Matthiessen
and
John A. Bateman

Pinter Publishers, London

First published in Great Britain in 1991 by
Pinter Publishers Limited
25 Floral Street, London WC2E 9DS

British Library Cataloguing in Publication Data

A CIP catalogue record for this book is available from the British Library
ISBN 0 86187 711 X

Typeset by Mayhew Typesetting, Rhayader, Powys
Printed and bound in Great Britain by SRP Ltd., Exeter

Contents

List of figures

Foreword

It gives me a very special pleasure to write this Foreword, because I have no less than three reasons for being delighted that this important and impressive work has now appeared.

This book gives the first full account of the development of the NIGEL grammar and the Penman framework for generation — first for English and then for Japanese. The NIGEL grammar has been described as 'the largest systemic grammar and possibly the largest machine grammar of any kind' (McDonald, Vaughan and Pustejovsky (1987:179)). It therefore goes far beyond the 'mini-grammars' (and even the 'midi-grammars') that are more typically used to exemplify the theoretical discussions in computational linguistics (CL). Indeed, it illustrates the way in which a new range of challenges arise when a rather fuller range of linguistic coverage is attempted. For its size and its richness alone, then, it must command respect. But it is far more than merely an account of a large grammar; the NIGEL framework includes, as well as the grammar itself, components that control the grammar, including an 'inquiry semantics'.

***

At this point it is right to pay a special tribute to the crucial role in the development of the NIGEL framework played by William Mann. Mann was by background a computer scientist with no special training in linguistics, but it was he who had the double insight to see, back in the late 1970s, the potential value for the project in text generation that he was planning of (1) giving the theoretical initiative, as it were, to the linguistics rather than to the computing, and (2) within linguistics, selecting a theory little in favour with computer scientists at the time (with the notable exception of the pioneering work by Winograd (1972) and Davey (1978)): Halliday's systemic functional grammar (SFG).[1]

In due course, Mann was joined at the Information Sciences Institute (ISI) of the University of Southern California by Matthiessen, and, working

within the overall framework initiated by Mann, Matthiessen gradually took on more and more responsibility for the detailed development of the NIGEL grammar and the semantic components above it. Other systemic linguists (including myself for a brief period) have contributed fragments, but above all Halliday himself has been a regular visiting consultant throughout the development of NIGEL.[2] Later Bateman, working at the University of Kyoto, adapted the original NIGEL framework for Japanese (Bateman, 1986), and later still he replaced Matthiessen at ISI when Matthiessen moved to the University of Sydney. In all of these complex variations of personnel — and many others not mentioned here — Mann played a crucial unifying and integrating role.

But as well as steering the Penman project successfully throughout the 1980s, Mann has made a significant theoretical contribution, with Sandy Thompson, by developing rhetorical structure theory. This is a model for monologue discourse which is having a considerable influence in CL circles, with which NIGEL interacts (as described in this book). Thus Mann's contribution to the work reported here — whether in terms of the original conception of the project, its overall structure, the work with Matthiessen on NIGEL in the early years, the year-in year-out direction of the project throughout the 1980s, or his particular theoretical contribution at the level of discourse structure — has been of fundamental importance. Without William Mann there would have been no NIGEL.

***

The first of the three roles in which I welcome this book is my role as the co-editor of two books already published by Pinter, in the *Open Linguistics* series. These are *New Developments in Systemic Linguistics, Volume 1: Theory and Description* (co-edited with M.A.K. Halliday, and published in 1987) and *Volume 2: Theory and Application* (co-edited with David J. Young, and published in 1988). Together, the two books are intended to give an overview of recent work in systemic linguistics.

The particular purpose of the second volume is to illustrate the fruitful interplay between theory and application in the many areas in which systemic linguistics has been applied. But one reviewer of that volume, while generally welcoming it, chastised the editors for not including a chapter representing the important field of computational linguistics. In fact, as was explained in the Foreword, we had originally intended to have just such a contribution — indeed, to have two: one by each of Matthiessen and Bateman. Now the full truth about that double omission can be told. The fact is that when Matthiessen's contribution finally came in — which was just as the shape of the book as a whole was being finalized — it turned out to be about *thirty thousand* words in length — itself approaching half

the length of a substantial book! I was immediately struck by both the quality of the account given there of the NIGEL framework and the impossibility of including it in the intended volume (unless we cut out several of the existing valuable contributions). Thus was born the idea of the present book. I asked Christian Matthiessen and John Bateman to consider joining forces to provide the definitive account of NIGEL, set as it is within the overall framework of the Penman Project. As I said in the Foreword to Fawcett and Young, 1988 (p. x), the present work 'should be thought of as an integral part of *New Developments in Systemic Linguistics*' because 'no picture of the theory and its applications would be complete without a chapter reporting work on the NIGEL grammar.'

Thus the present volume, while belonging squarely in the series *Communication in Artificial Intelligence*, also belongs in principle in the *Open Linguistics* series, as a book-length example of the fruitful interplay between theory and application in the field of computational linguistics. But note that, despite the length of treatment allowed, this book is still not a full account of current work in this field, as we shall see below; this is essentially the account of one major project in this field that was developed and complemented at two sites: the NIGEL framework.

***

Second, I welcome this significant book as a fellow researcher in this field. In my role as Director of the COMMUNAL Project, which is also at heart a text generation project and which also uses systemic functional grammar, I pay tribute to the courage, imagination, intelligence and sheer hard work with which the researchers on NIGEL and Penman attempted — and succeeded at — their enormous task. By so doing they made it easier for others to attempt to build more or less related models. These include Bateman's Japanese model, Patten's SLANG (Patten, 1988), the GENESYS component in the COMMUNAL Project at Cardiff (Fawcett, 1988; Fawcett and Tucker, 1990; Fawcett, 1990; and Tucker, to appear), and two recently begun projects: the KOMET Project at Darmstadt of Steiner, Bateman and colleagues (Steiner *et al.*, 1990; Wanner and Bateman, 1990), and even more recently started work by Matthiessen and colleagues at Sydney.

Let me make explicit one particular aspect of the debt owed to the pioneering work of Mann and Matthiessen by the rest of us. We have been like cross-country skiers following the tracks of a trail blazer, with the way (partially!) smoothed before us. The key point is that the task of implementing SFG computationally was very different from — and much greater than — that of implementing the current post-Chomskyan models, e.g. in parsers. The reason is that relatively little work had been done on turning SFG,

which had been developed primarily for the hand analysis of texts, into a fully explicit, generative model. The effect is that a great debt is owed to the NIGEL team by those of us who have come after — whether we have been building explicitly on the base of the NIGEL framework for work on another language (as at Kyoto and Darmstadt) or exploring alternative ways of controlling the grammar (as with Patten at Edinburgh) or developing both a new framework and an alternative SFG (as at Cardiff). The great advantage shared by us all has been that we have known that, in principle, *the task is doable*.[3] It is this single fact, in my view, that has been the biggest contribution to giving us later workers in this new field of research the confidence to keep going on those occasions when, if we had not known that others had solved similar problems in a similar framework, we might have been overfaced (to use a good North Country word) by various problems of implementation. For us at Cardiff it was not that we borrowed solutions to the problems from NIGEL (because we did not, believing in the value of seeking our own solutions and so enriching the available experience) — but that we knew that it could be done.

***

I wrote in the Foreword to the *New Developments in Systemic Linguistics: Theory and Description* (Halliday and Fawcett, 1987):

> The very notion of developing a theory and then applying it is — or should be — a nonsense. The fact is that theories develop most creatively when descriptions of language that 'realize' them are tested in applications of various sorts — and, very often, the theories are stretched by the exercise and in due course reworked. It might therefore have been . . . appropriate . . . if we had given the second volume some such title as 'the theory-description-use-theory cycle'.

The position taken by Matthiessen and Bateman is precisely the same. As they say (Chapter 1, Section 1.1): 'Application is in fact an opportunity to work on theory'. Thus, this work is not 'merely' the application of a complex theory of language to an equally complex and challenging task (as if any application were 'mere'); it is equally a contribution to linguistic theory.

In particular, it is a major contribution to the further development of systemic linguistics. But newcomers to this theory should not be daunted; the book is written in such a way that you will be led gently into an understanding of the theory — and in due course an understanding of some of the issues that arise when building computational models of SFG: issues both for the implementation and for the theory.

Thirdly, then, I welcome this book in my role as a student of language and a fellow explorer of the systemic functional route to a better understanding of this all-pervasive and ever-fascinating phenomenon. This work

should interest all linguists, irrespective of whether they happen to believe that theory should be tested and developed in the research paradigm of the computer, as it is in this case.[4]

There seems to be a view among linguists working in non-systemic frameworks that there is essentially just one systemic linguistics, i.e. that of Halliday (e.g. 1985). This is fairly far from the current reality. The fact is that Halliday does not lay down a currently 'correct' version to which all true believers should adhere. Indeed, he himself often offers alternative positions (e.g. on 'transitivity' in Halliday, 1985). He has long recognized the 'ineffability of language' (Halliday, 1984), and tolerates and encourages the exploration of alternative approaches within the general systemic framework. Thus the NIGEL grammar as developed by Matthiessen (and, as stated above, by many others who have visited ISI and worked on it) is in some respects different from the grammar found in Halliday, 1985, and the GENESYS lexicogrammar and generator in the COMMUNAL Project at Cardiff is different in many significant ways from both. Similarly, new ideas potentially significant for SFG theory are also appearing in the KOMET Project at Darmstadt (e.g. Wanner and Bateman, 1990).

Others working in the systemic framework (but not necessarily in the computational paradigm) have developed yet other ideas, some published (e.g. many of the contributions in Halliday and Fawcett, 1987; Fawcett and Young, 1988, and Davies and Ravelli, to appear). But many interesting ideas are simply presented at the annual International Systemic Congresses or discussed in local groups, without being formally published. Sometimes accounts of these ideas find their way into *Network*, the newsletter for those who want to keep up with news of the systemic linguistics community, but more often, regrettably, they do not.[5]

There is almost as much difference between some of these systemic approaches (e.g. those of Halliday, Fawcett, Gregory and Martin) as there is between members of the broad family of 'post-transformationalist' theories, such as generalized phrase structure grammar (GPSG), lexical functional grammar (LFG) and government and binding (GB). The importance of the present book in this respect is that it sets out the case for doing things in the way that they have been done in the NIGEL framework more clearly than is usually done in SFG, and that it discusses — though inevitably briefly — at least some of the alternatives. So yet another service that this excellent book performs is to provide a beginning — in a way that Halliday's key text *Introduction to Functional Grammar* does not (and, given its purpose, could not) — to the serious discussion that is now needed of the many questions that arise within systemic linguistics. Typically, these do not get discussed in depth in a forum more public than, for example, the research seminars held to thrash out current problems in the

COMMUNAL Project — and in similar semi-formal groupings, wherever systemic linguistics are at work.

The 'Introduction' to Halliday and Fawcett (1987) sets out eight 'new developments' (or, in another terminology, 'issues') within systemic *linguistic theory*. In building the GENESYS sentence generator in the COMMUNAL Project we have deliberately set out to broaden the collective SFG *computational* experience in two ways. First, we have chosen to explore alternative routes to solutions to many of the same problems that arose in developing NIGEL (e.g. by using the declarative language of Prolog rather than the procedural language of Lisp); by using a different but equally well-tested SFG description of English in which the networks are explicitly at the level of semantics (based on Fawcett, 1980); by giving central status to the selection expressions as a semantic representation; and by attempting to design the realization rules so that they are intrinsically reversible (see O'Donoghue, 1991). Second, we have additionally built, or begun to build, components that would in principle have been built for NIGEL if there had been time, and which will no doubt be developed in due course (e.g. the realization of meaning in intonation (Fawcett, 1990); the integration of meanings realized lexically in the same network as that for meanings realized grammatically and intonationally (Fawcett and Tucker, 1990; Tucker, to appear); the use of probabilities and preferences in system networks (Fawcett, to appear a and b); discourse modelled as social interaction (Fawcett, van der Mije and van Wissen, 1988). All of these innovations, like those in NIGEL, raise questions of principle and should contribute, in the spirit of the present book, to future discussions of systemic theory.

Here, in conclusion, is a selective checklist of questions that are raised in the context of work in the computational paradigm, and on which a SFG linguist or computational linguist should have (or be working towards having) a position. Virtually all are as relevant to those who want a good description of language for, let us say, describing a text of literary or social interest as they are to building better models for text generation or understanding. The hope of all those working in the demanding and exciting field of CL is that the various pressures on us to deliver working systems — pressures that our confidence in systemic theory enables us to feel well able to meet, given time and adequate resources — will not prevent us from somehow finding more time than we have had in the past to give to considered discussion, comparison and experimentation in relation to these matters.

1. Can the system networks in the lexicogrammar be semanticized further than they are in NIGEL, without overstraining the realization rules (or 'statements')?

2. Given that 'realization' rules/statements are inherently 'actualization' rules/statements, are they also genuinely 'realization' (i.e. inter-stratal) rules/statements?
3. Is it necessary to postulate a higher stratum of even more 'semantic' choices, still within the semiotic system of language, as opposed to their being part of the 'belief system' or 'upper model' (assuming the rough equivalence of these terms)?
4. What is the role of probabilities in (a) generation and (b) understanding?
5. At what level of planning is it most appropriate to handle register and other types of variation that are open to users of a language?
6. Should there be, in a SFG approach, a role for anything remotely like the standard notion of a 'lexicon' that is separate from the 'grammar' (e.g. in parsing)?
7. What is the relationship of the traditional linguistic concept of a 'description' of a language to the computational concept of language as a processing device for turning (very roughly) 'meanings' into 'sounds' and back? Are there 'neutral' descriptive models?
8. Should models of producing text be derived from models of understanding text, or *vice versa*, or from 'neutral' descriptive models?
9. What is the role of 'gates' (roughly, one-feature systems)? Are all gates essentially of the same type, or are some part of the system network and some part of the realization apparatus?
10. What is the full range of possible ways of handling the inter-stratal relationship of realization (which has been much less fully discussed so far than networks), and what is the optimal apparatus and notation?
11. What should the criteria be for deciding between alternative approaches within SFG to phenomena such as the 'clause complex' (e.g. 'hypotaxis' vs. 'embedding')?
12. The final question to be raised here is the one that is perhaps the largest of all for linguistics. If a SFG generator, with its system networks and realization rules, is the 'sentence planning' stage of text planning (i.e. if it is the last of possibly three broad stages of planning a text), then how far is linguistics an autonomous discipline? Just as most systemicists (and many others) would now reject the Chomskyan notion of 'autonomous syntax', claiming that 'meaning potential' (or 'semantics') has to be given its due place in the picture, so too it may be misleading to try too hard to produce 'autonomous linguistics' models — in the sense of models that generate, from a base that specifies the meaning potential (as in SFG), all and only the grammatical strings of words for a language. Any adequate model must find ways of handling the various constraints that are needed *at the*

appropriate level. But are all of these 'levels' within the semiotic system of language itself? In other words, is the concept of 'generation', in its formal language theory sense rather than its natural processing sense, still exerting an undue influence on the field?

Matthiessen and Bateman touch on almost all of these issues — and indeed on many others — and on some they have a clearly stated position. When they take a position it must be treated very seriously, because it has been tested in the fire (and sheer hard slogging work) of a computational implementation. It is for their contribution to the developing discussion of these and other crucial questions of linguistic theory that I personally welcome this book most of all. And for other readers, such as those who are simply seeking a clear exposition of how a SFG can be used in text generation, the book will be equally valuable.

Robin P. Fawcett
Radyr, June 1991

Notes

1. Most researchers in parsing have preferred to work with broadly Chomskyan models of language that are explicitly influenced by the concepts and procedures of formal language theory; this is despite Winograd's seminal work on natural language understanding, which used systemic grammar (Winograd, 1972) and Winograd's continuing use of it, e.g. as the basis for the linguistic descriptions used in his widely used textbook *Language as Cognitive Process* (1983).
2. Indeed, the name 'Nigel' is itself an indication that Halliday is the 'father' of the grammar, in that 'Nigel' is the name used for Halliday's son in Halliday's celebrated account of Nigel's language development in *Learning How to Mean* (Halliday, 1975).
3. Winograd, 1972 and Davey, 1978 were earlier breakthroughs, but it was the sheer size and range of coverage of NIGEL that gave the additional confidence.
4. Nonetheless, it is significant that the testing and development *have* been done in this way; computational implementation has in the last decade — and for good reasons, I believe — become one of the leading ways of testing the validity of a linguistic theory. This should be alongside, as always (1) the time-honoured but ultimately unverifiable test of whether or not descriptions of actual texts based on a detailed description of a language give the analyser a sense of insightfulness, and (2) the potentially valuable but in practice still unreliable craft of psycholinguistic testing.
5. Those of us working in higher education these days are almost all too busy with teaching and administration to give as much time as we should to writing — and, indeed, all too often, to reading. Those with research posts may similarly have pressures on them to produce working models that cut down drastically the time that is left for interaction with co-researchers (whether orally, or through the writing and reading of papers) and too on the time for simply letting ideas develop. Research students: make the most of your few golden years; they may never come again (or only with a great struggle)!

Bibliography

Bateman, J.A. *Text Planning for a Systemic Functional Grammar of Japanese*. Technical Report. Kyoto, Japan: Department of Electrical Engineering, University of Kyoto, 1986.

Berry, M.M., Butler, C.S. and Fawcett, R.P. (editors) *Meaning and Choice in Language: Studies for Michael Halliday. Volume 2 Grammatical Structure: a Functional Interpretation*. Newark, N.J.: Ablex, to appear.

Davey, A. *Discourse Production: a Computer Model of Some Aspects of a Speaker*. Edinburgh: Edinburgh University Press, 1978.

Davies, M. and Ravelli, L. (editors) *Advances in Systemic Linguistics: Recent Theory and Practice*. London: Pinter, to appear.

Fawcett, R.P. 'Language generation as choice in social interaction'. In Zock and Sabah (editors) 1988b, 27–49, 1988.

——— 'The computer generation of speech with semantically and discoursally motivated intonation'. In *Procs of 5th International Workshop on Natural Language Generation*, Pittsburgh, 1990.

——— 'A systemic functional approach to selectional restrictions, roles and semantic preferences', to appear a. Accepted for *Machine Translation*.

——— 'A systemic functional approach to complementation', to appear b. In Berry, Butler and Fawcett, to appear.

Fawcett, R.P. and Tucker, G.H. 'Demonstration of GENESYS: a very large, semantically based systemic functional grammar'. In *Procs of 13th International Conference on Computational Linguistics*, Helsinki, Volume 1 pp. 47–9, 1990.

Fawcett, R.P. and Young, D.J. (editors) *New Developments in Systemic Linguistics, Volume 2: Theory and Application*. London: Pinter, 1988.

Halliday, M.A.K. *Learning How to Mean*. London: Arnold, 1975.

——— 'On the ineffability of grammatical categories', in Manning, Martin and Mc Calla 1984. pp. 3–18, 1984.

——— *An Introduction to Functional Grammar*. London: Arnold, 1985.

Halliday, M.A.K. and Fawcett, R.P. (editors) *New Developments in Systemic Linguistics, Volume 1: Theory and Description*. London: Frances Pinter, 1987.

Kempen, G. (editor) *Natural Language Generation*. Dordrecht: Martinus Nijhoff, 1987.

McDonald, D.D., Vaughan, M.M. and Pustejovsky, J.D. 'Factors contributing to efficiency in natural language generation'. In Kempen 1987, pp. 159–81, 1987.

Manning, A., Martin, P. and Mc Calla, K. *The Tenth LACUS Forum 1983*. Columbia: Hornbeam Press, 1984.

O'Donoghue, T.F. 'A semantic interpreter for systemic grammars', In *Procs ACL Workshop on Reversible Grammars*. Berkeley, California, 1991.

Patten, T. *Systemic Text Generation as Problem Solving*. Cambridge: Cambridge University Press, 1988.

Steiner, E.H., Bateman, J.A., Maier, E., Teich, E. and Wanner, L. *KOMET: Department Plan*. Technical report. Darmstadt, Germany: GMD/Institut fur Integrierte Informations- und Publikationssysteme, 1990.

Tucker, G.H. 'Cultural classification and system networks: a systemic functional approach to lexical semantics', to appear. In Berry, Butler and Fawcett, to appear.

Wanner, L. and Bateman, J. 'Lexical cooccurence relations in text generation'. In *Procs of 5th International Workshop on Natural Language Generation*, Pittsburgh, pp. 31–8, 1990.

Winograd, T. *Understanding Natural Language*. Edinburgh: Edinburgh University Press, 1972.
——— *Language as Cognitive Process. Volume 1: Syntax*. Reading, Mass: Addison-Wesley, 1983.
Zock, M. and Sabah, G. (editors) *Advances in Natural Language Generation Vol. 2*. London: Pinter, 1988.

Acknowledgements

The research discussed in this book has been carried out throughout the 1980s in several places: we are greatly indebted to the many people who have contributed in various ways over the years. Without the insight of Bill Mann the Penman system would not have come into existence: he conceived of the system, involved us, and directed the Penman project for most of the 1980s. We have benefited from our many discussions with him — just as with the other project members and co-workers at ISI: Robert Albano, Susanna Cumming, Ed Hovy, Bob Kasper, Johanna Moore, Cécile Paris, Lynn Poulton, Norman Sondheimer, Yu-Wen Tung, Richard Whitney and others. To Yu-Wen we also owe a special debt for his research on parallel generation, which forms the basis of our Chapter 10, and the work presented in Section 12.5 is joint work with Cécile Paris. Furthermore, several visitors and consultants have been crucial: Robin Fawcett, Peter Fries, Michael Halliday, Ruqaiya Hasan, Jim Martin, Sandy Thompson and others. Kary Lau at ISI receives a special mention and thanks for her work in preparing the final versions of most of the figures for the book.

The research contexts at Kyoto University and GMD/IPSI in Darmstadt have been very significant in expanding and supplementing the work at ISI. Thanks are due here to Professors Makoto Nagao and Jun-ichi Tsujii for being very supportive of the development of our prototype treatments of Japanese at Kyoto, and to the eager and thorough participation of the students, particularly Gen-ichiro Kikui, Atsushi Tabuchi and Li Hang; John Bateman offers further special thanks to Wendy (Jones) Nakanishi for prompting his foray into Japanese linguistics.

As will be abundantly clear from the book, our fundamental theoretical debt is to Michael Halliday. He has also provided constant support and encouragement in many exchanges around the globe. Bill Mann, Paul Schachter, Sandy Thompson and Mayumi Masuko commented on earlier materials on which we have drawn in preparing this book, and Cécile Paris and Erich Steiner have both made many helpful and detailed comments on

the book drafts themselves. Robin Fawcett suggested that we should transform a book chapter into a book and so he is responsible for the existence of the present work! Through his own closely related research, he has been a constant source of inspiration, always willing to help and challenge.

As the above begins to suggest, it would almost be easier to list the places, countries, and continents where parts of this book have *not* been written, either by one of us individually or by both of us together, than those where they have been. This has naturally presented various technical and organizational problems and so, finally, we would like to thank Pinter Publishers, particularly Vanessa Couchman, Heather Bliss and Nicola Viinikka, for continued patience in the face of ever postponed delivery dates!

1 Introduction: text generation as an application of linguistic theory

The nineteenth century saw the culmination of the Industrial Revolution in the Western World. The revolution was a material one in many respects and the sciences that came of age at the time were the sciences of matter — inanimate and animate. Physics has accordingly emerged as the paradigm for doing science and this further foregrounds the material interpretation of our world. However, there are signs in modern physics that we need another kind of interpretation, a semiotic one where information plays a major role. This is not a new way of looking at the world, of course. The question of how to manage and represent information was as important in the sixteenth and seventeenth centuries when the modern scientific period started in the West and was in its predominantly taxonomic phase as it is now. But this complementary semiotic interpretation of our world has only recently come into prominence.

One central reason for this is that we now have tools for studying and manipulating information that are likely to be as important as the telescope, the microscope, and the steam engine were for the development of the scientific material world view — tools such as the tape recorder and the computer. These tools allow us to record and disseminate information in ways that were not made possible by the two previous milestones in the manipulation of information, the inventions of writing and printing.

The most elaborate and sophisticated information system we know is natural language and that is good reason to focus on disciplines that provide insight into language. There are several, of course, reflecting the various levels at which language exists (social, psychological, neurological, aesthetic, and so on). Indeed, this multiplicity is likely to prove one of the focal points in the breakdown in the traditional disciplinary boundaries and to be important in the foundation of a more thematic approach to the exploration of our world. It already figures prominently in two major interdisciplinary developments — cognitive science and semiotics.

1.1 Text generation

This book lies at the intersection of two paradigms for working on and with language: it is concerned with the exchange between linguistics and artificial intelligence (AI) in the context of TEXT GENERATION. Text generation can be characterized as one mode of information processing: information that is stored at a higher order of abstraction than wordings (grammatical structures and lexical terms) is organized and re-expressed over a number of steps so that it can be presented as a worded text. Together with text-understanding, parsing and machine translation, text generation is often known as natural language processing (NLP) and is pursued within computational linguistics.[1] This book deals with the application of linguistic theory and description to the task of text generation, drawing on AI techniques and methods.

Text generation research is motivated by the need to present information stored in or produced by the computer in a data base, an expert system or the like — information that would not otherwise be accessible to a person who is not a computer expert. Research on text generation thus usually involves getting the computer to function as a writer for some fairly restricted writing task (or occasionally as a speaker; cf. Sigurd, 1984; Davis and Hirschberg, 1988; Davis and Schmandt, 1989): given a need for text, the computer will produce text automatically in response to this need. We can see text generation and text understanding as complementary processes for communicating with the computer, i.e., exchanging information with it, in our own terms — those of natural language. Looked at in this way they represent a natural extension of the history of programming languages, which started out as machine languages, but have developed more and more to accommodate human principles for organizing information.

One kind of reason, then, why one might want to build a text generator would be to automate the production of weather forecasts by building a weather forecast generator. The generation process would involve finding the relevant information about yesterday's, today's and tomorrow's weather, organizing this information according to the format of weather forecasts, and then wording this information by means of grammatical structures and lexical items (cf. Isabelle, 1984; Kittredge, 1987). We will see many other applications of text generation of this kind below.

Another kind of reason for constructing generators comes from within linguistics. It is increasingly being realized among those working on text generation in a purely computational context, i.e., at departments of computer science or electrical engineering, research institutes concerned with computational research, and within the computer industry, that it is important to be aware of text generation as an application of linguistic

theories and principles (we motivate this position more in Chapter 3). However, this has not yet been so well accepted among the linguistic community. As an application of linguistic theory, text generation stands in the same relationship to theory and description as do, for example parsing, pedagogic grammar, contrastive analysis, typological studies, the quest for linguistic universals, and literary studies. Linguistic theory is responsible to all applications such as these and they all involve theory revision and extension. Application is in fact an opportunity to work on theory. Thus, it is necessary to appreciate the value of *text synthesis* as a complementary approach to text analysis in the study of text and the other semiotic systems instantiated in text.

This has already motivated a number of linguistic researchers to work on issues related to the creation of text — with or without computers: Bengt Sigurd's (e.g. 1984) COMMENTATOR system developed at Lund University is a good example of the former, and Fawcett (1973, 1980), Halliday (1973, 1977), Chafe (1977, 1979), and Dik (1987) are good examples of the latter;[2] also, current work at Sydney University exemplifies both kinds of work. This tendency is certain to increase with the ready availability of ever more powerful, and yet cheaper, computers: linguistics departments will no longer be unable to afford machines capable of running the sophisticated software necessary for meaningful linguistic research. Text synthesis could then become just as valuable to discourse study as speech synthesis has become to phonetics. But the initiative for this needs to come from within linguistics, which is why it is important for linguistics also to conceive of text generation as a linguistic research task.

1.2 Systemic-functional linguistics

The particular approach to language around which we will centre our account of the application of linguistic theory to text generation is SYSTEMIC LINGUISTICS, which originated in Britain but is now used and developed internationally. Systemic linguistics started with M.A.K. Halliday's work in the early 1960s and has been developed by him and others into a holistic theory of *language in context*. Halliday's immediate linguistic environment was that created by Firth in the 1930s through 1950s.[3] One important theme was text in context, a way into language emphasized by the anthropologist Malinowski, who influenced Firth and also later systemic linguists. Halliday has also been influenced by, e.g., Prague School linguists, Hjelmslev, and Whorf.

In the systemic interpretation, language is seen as a resource for making meaning. This resource is organized functionally according to a small

number of simultaneous highly generalized functions (metafunctions), concerned with the representation of experience (ideational), symbolic interaction between speaker and listener (interpersonal), and the presentation of information as text in context (textual). This resource is also stratified into subsystems forming three levels of abstraction ((discourse) semantics, lexicogrammar and phonology/graphology). Lexicogrammar is the unified resource of grammar (syntax and morphology)[4] and lexis (vocabulary). Grammar and lexis are simply different aspects of the same system; both are simultaneously ideational, interpersonal, and textual. For instance, ideational lexis is the (taxonomic) organization of vocabulary as denotation, interpersonal lexis is the connotative organization, and textual lexis is concerned with the use of lexical resources in the creation of cohesive text.

In systemic linguistics, each level of the linguistic system is organized as a network of interrelated choice points (a system network), making explicit what resources are available. Selections from this network are realized (expressed, coded) by structures, items, etc. Choice is thus the primary principle of organization, not structure. The linguistic system as a whole is furthermore 'embedded' in context, which can be interpreted as higher-level semiotic systems (e.g., systems of social relationships and ideological systems).

The reasons for our choice of systemic linguistics as the theoretical linguistic foundation in text generation will emerge below (particularly in Chapter 5), but it is important to note in the present context that systemic linguistics differs sharply from Chomskyan linguistics, with Government and Binding theory as the current principal manifestation (Chomsky, 1981), and the various frameworks that have been presented as alternative ways of answering the questions about language that Chomsky is concerned with (Lexical Functional Grammar — LFG: Bresnan and Kaplan, 1982; Generalized Phrase Structure Grammar — GPSG: Gazdar, Klein, Pullum and Sag, 1985; and so on). First of all, systemic linguistics interprets and represents language not as a rule-system for generating structures but as a *resource* for expressing and making meanings. The interpretation of language as a resource makes a fundamental difference when we build a text generation system since the purpose of such a system is precisely to express and make meanings. Second, systemic linguistics has expanded the boundaries of linguistics and made them more permeable from outside. It can thus help facilitate the interdisciplinary exchange we are concerned with.

These and other related differences should be emphasized in the present context since the image of linguistics in computational areas is often the Chomskyan one. This is probably one significant factor in the occasional

wholesale rejections of linguistics that may be observed within Natural Language Processing, which is unfortunate since there are other ways of doing linguistics, other ways of conceiving of, and theorizing about, language. Even when the Chomskyan contributions, particularly in the area of syntax, are accepted, the systemic alternative (as well as others, of course) is worth considering. Many areas which are crucial to computational linguistics (functional grammar, discourse, context, register, and so on) simply fall outside Chomskyan theorizing but are part of the systemic domain. We will illustrate many of these areas explicitly in the chapters below: thus, our choice of systemic linguistics as the theoretical foundation for our work on text generation is motivated both because it addresses issues that are crucial to natural language generation and because it facilitates the interdisciplinary exchange we believe is now necessary.

1.3 Organization of the book

We have divided our discussion into four parts to show the progression from the opening up of the exchange between work on text generation and systemic linguistics through particular research applications and conclusions that can be drawn from them so far to future directions building on these conclusions. The specific context of our discussion is provided by two particular text generation systems with which we have been involved: the Penman system for the generation of English text and the Kyoto system for the generation of Japanese text.

Part I, Text generation and systemic linguistics: opening the exchange (Chapters 2–4), introduces text generation and systemic linguistics. Chapter 2 deals with the model of text generation in natural language processing and Chapter 3 gives a brief interpretative survey of past work on text generation within a computational setting. Chapter 4 takes up the issue of the exchange between linguistics and text generation, indicating what the division of labour has been up to now.

Part II, Systemic linguistics and text generation: the basic theoretical framework and two computational instantiations (Chapters 5–7), lays the linguistic foundation for the rest of the discussion and introduces two research applications. Chapter 5 is a very brief synopsis of those aspects of systemic theory that are particularly relevant to text generation. Chapter 6 presents this framework as it is embodied in the two research applications, focusing on the level of grammar. Chapter 7 draws on the grammars of Japanese and English to illustrate the use of the framework.

Part III, Up to and beyond the limits of the basic framework (Chapters 8 and 9), explores the question of how far the current computational

modelling of systemic theory has taken us and where we have to return to the theory for insights into how to revise the model. We divide this discussion along two dimensions of organization drawn from systemic theory: Chapter 8 discusses extensions within grammar in terms of the METAFUNCTIONAL decomposition of grammatical resources, and Chapter 9 discusses extensions that go beyond the grammar to higher STRATA of the linguistic system.

Part IV, Future directions for computational systemic-functional linguistics (Chapters 10–12), takes up a number of issues that are only beginning to be explored and that are thus one step beyond those discussed in Parts II and III in the evolving exchange between theory and application. Chapter 10 addresses the possibility of parallelism in text generation based on multi-functional systemic theory and Chapter 11 turns to the task of developing dynamic interpretations and representations of the linguistic system. Chapter 12 explores the relationship between the general and the specific in text generation — the tension between the generalized notion of text generation and the specific generators created for generating reports, encyclopaedic entries, etc. — in terms of the systemic concepts of context and register variation.

Finally, to conclude the book, Chapter 13 briefly discusses other related systemic-functional work that is currently under way and which we feel will be important for the future of computational linguistics, systemic linguistics, and their interaction. We organize this in terms of the dimensions of components of the theoretical framework as developed throughout the book — this should be of use to readers who wish to follow up on particular aspects of the theory and practice of systemic-functional linguistics in general and its application to computational natural language processing in particular. Further collections of references are provided in supplemental tables following Chapters 2, 3 and 5 and a general glossary of terms precedes the bibliography.

Notes

1. Natural language processing has not previously been taken to include work on various kinds of data processing where the computer is used as a research tool to examine and discover statistical patternings. However, more recent results in the analysis of large-scale corpora are finding their way into natural language processing proper.
2. There are also examples within psycho-linguistics; but they tend to be further away from the problems we are concerned with here.
3. It has to be emphasized that the Firthian environment was fundamentally different from the American linguistic environment in which Chomsky grew up and reacted against. Many of the issues and insights that have only appeared on the agenda in the late 1970s and 1980s as a reaction against the version of generative linguistics developed in the 1960s were already on Firth's agenda (see e.g., Henderson, 1987).
4. Note that the term grammar is thus used in its traditional sense and not in its more recent Chomskyan sense of '(model of) the linguistic system', where it can come to include not only syntax and morphology but also semantics and phonology.

Part I Text generation and systemic linguistics: opening the exchange

2 The model of text generation in natural language processing

As we will see in our historical overview of text-generation systems in Chapter 3, such systems vary quite significantly in basic design. In this chapter, we discuss briefly what a text generation system of the early 1990s looks like when conceived in computational linguistic terms and identify a number of key components, presenting them in terms of one design.

2.1 The design of a representative generation system

In the design of a text generator, one general tendency is to distinguish between *resources* such as knowledge resources, a user model, rhetorical resources, grammatical resources, and lexical resources; and *processes* or procedures that employ these resources. Examples of processes include knowledge selection, text planning (organization), lexicogrammatical expression, and editing. The general process of generating text is often seen as *goal pursuit* — a powerful metaphor developed within AI which defines processes that may be performed in order to achieve as yet unrealized but desirable states. In the particular case of text generation, therefore, *communicative goals* are specified and the generator produces a text in order to attain these goals.[1]

A simple design is diagrammed in Figure 2.1. For concreteness, we will centre our opening description on Mann's (1982) design plan for the Penman generator, which, as one of the broadest scoped generators that has been developed, offers an orienting framework for most of the features we need to introduce. It has also provided the starting point for a number of current developments in text generation (e.g., Matthiessen, 1989; Steiner *et al.*, 1990a), text analysis (Kasper, 1989b), and machine translation (Bateman, 1989b; Hovy, 1989a) and so continues to be relevant for current research efforts.

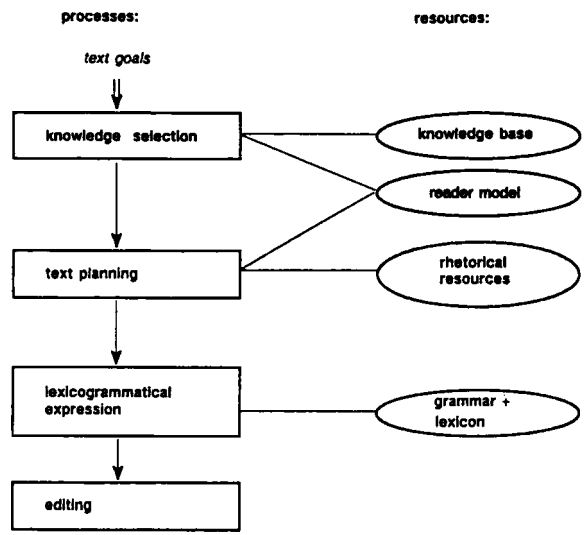

Figure 2.1 Simple sequential generator design

The generator starts by invoking a subprocess, KNOWLEDGE SELECTION (also called SEARCH), which extracts the relevant knowledge from the knowledge resources of the generator. The selection process can be guided by a model of the intended user, the user model (also sometimes called the reader or hearer model), so as to avoid including information already known to him or her and to include background information not known to him or her.[2]

Text PLANNING then arranges the extracted knowledge into a RHETORICAL STRUCTURE using the strategies offered by the rhetorical resources, again consulting the user model if available. In Penman, for example, the overall organization of the text is planned in terms of *Rhetorical Structure Theory*. The result is a hierarchy of rhetorical units, each terminal unit typically corresponding to a non-embedded clause.

The process of LEXICOGRAMMATICAL EXPRESSION then takes over and executes the text plan by turning it into wordings, relying on the lexicogrammatical resources. In Penman, for example, the process of lexicogrammatical expression uses the *Nigel* component of the system, which is organized into two levels (strata): a systemic-functional grammar and a semantic interface