Computers and the Humanities 16 (1982) 119–129
North-Holland Publishing Company

BOOK REVIEWS

Proceedings of Apollo Agonistes; The Humanities in a Computerized World, M.E. Grenander, ed., The Institute for Humanistic Studies, State University of New York at Albany: 1980; 2 volumes, xxxii + 438 pp.

By Susan Hockey

A conference called 'Apollo Agonistes'? An unusual title for a conference. These volumes of proceedings suggest it was an unusual conference. The introduction points out that the call for papers was sent to 12 000 persons primarily chairmen and deans of selected departments but also directors and chancellors of foreign universities. Not many of the 12 000 elected to attend. The list of participants at the end of Volume 2 gives some 150 names of which less than 20 were from outside the United States.

The introduction states that "the Proceedings are a record of the actual symposium". This is so, for besides a large number of papers of which more later, we also have the complete programme and even the menu at the symposium banquet where we learn that participants sat down to a meal of chilled vichyssoise, chicken supreme, salad and pêche melba, accompanied by New York State Pirot Chardonnay. The only item lacking from the proceedings is the text of the keynote address which followed the banquet. However, not to worry, we are told that we can hire or purchase a videotape of it if we wish.

What of the papers? With a sub-title of 'The Humanities in a Computerized World', we can expect almost anything and almost anything we get, but to an outsider the dominant theme is eccentricity. What can one make of Sam N. Lehman-Wilzig, 'Frankenstein redux: The computer-criminal of the Lyborg age' in which he dwells on the legal implications of a robot turning into a Frankenstein. This paper is not helped by its footnotes, or lack of them. Nine pages are liberally scattered with footnote references (59 in all), but only footnote 1 exists – it announces that a footnote sheet will be distributed at the conference or by request to the author.

Credibility is hardly gained by reading through the other papers. What can one make of 'The caduceus and the lyre: Forms of cognition' or 'Talking statues in Elizabethan and Stuart England' or 'Leibniz: Precursor of the computerized world'? On delving into the volumes one can find equally entertaining quotes. "One possible solution to this *aporiae* is to advocate a complementarity of approaches. The polarities of objective–subject, scientistic–humanistic, naturalistic–hermeneutic have led to little constructive synthesis" state McCluskey et al. in conclusion to their paper on 'Prediction, power, and control: Games, simulation techniques and systems theory in the social and behavioural sciences'. Crosjean on 'Biblical perspectives on the use of number in the computer age' concluded "Numbers serve the needs of society at a deeper level through symbolic development". A British onlooker may well be amused at Plank's conclusion, "The computer is the capitalization, the bureaucratization, the bourgeoisification of information and of the world as information – and as such is as American as apple pie"; this comes at the end of a paper on 'Information vs truth: The computer and the history of ideas'.

However, amongst the eccentrics, it is possible to find some papers of more relevance to *CHum* readers, but all on the subject of humanities com-

Susan Hockey teaches computing in the humanities at Oxford University Computing Service and is a Fellow of St Cross College, Oxford.

puting are rather superficial and have an air of not reaching the required level of nuttiness. Dibble gives a useful overview of automation in libraries from the librarian's viewpoint. Fabian discusses publication in microform and Sorensen's view of new computer applications in content analysis must be taken seriously. Sedelow, Holzman and Svartvik all discuss traditional text analysis computing problems, but none with sufficient depth to make their papers useful.

Also floated are other ideas which are less relevant to *CHum* readers but nonetheless serious. There are three useful papers on the legal and sociological problems encountered when computers are used in courts of law, bringing to mind a recent case in Britain in which evidence produced by a computer was held to be inadmissible. Parsons and Debenham's description of a computer model of rational decision-making would be more interesting if it gave some decisions which it had made. Scragg and Wilkes give a useful discussion of the psychological effects of computers on their users. There is not much to commend these volumes to *CHum* readers, other than entertainment value, but they do score very highly for nonsense quotes, sci fi and its attendant jargon. They are certainly not recommended reading for those whose interest in humanities computing is just awakened.

Archives and the Computer, Michael Cook, Butterworth, London: 1980; 152 pp., ISBN 0 408 10734 0.

By Meyer H. Fishbein

Most large national archives have computer-oriented systems for managing their holdings. Several have accessioned machine-readable records. During the last decade smaller archival institutions in developing nations, as well as regional and university archives in the developed nations, have installed systems or are studying potential uses of automation. Nevertheless, courses and textbooks devoted to archival automation have been absent until recently. Michael Cook, archivist of the University of Liverpool, is helping to overcome this neglect with his *Archives and the Computer*.

This text is designed primarily to provide archivists in the Third World, as well as students of archival theory and practice, with self-instruction on automation. It explains feasibility studies, hardware and software, retrieval systems, the use of automation for managing current and non-current records, and the selection of machine-readable records for permanent retention. Specific illustrative applications are described in England, the United States, and the Ivory Coast.

Cook succeeds in explaining complex operations in relatively simple terms. An appended glossary aids the reader over the hurdle of technical terminology. His explanation of input methods, visual display units and linkage with on-line systems are especially clear and useful to the uninitiated. Further editions may well deal in more detail with word processing and with the Italian adaptation of complete text retrieval for early records. For those who are unfamiliar with archival automation, the text is a useful introduction.

Meyer H. Fischbein is a consultant on archives automation in Bethesda, MD.

Tölvukönnun á tíðni orða og stafa í íslenskum texta (On Computer-Generated Word- and Letter-Frequencies in an Islandic Text), Baldur Jónsson, Björn Ellertsson and Sven Þ. Sigurðsson, University of Iceland, Reykjavík: 1980; 148 pp.

By Marina Mundt

The purpose of this report, according to the authors (p. 140), is threefold: to account for the execution of the project, i.e., the achievement of the first computer-made concordance and a variety of frequency studies at the graphic level of any modern Icelandic text of some length; to account for results, mainly in tabular form, concerning

Marina Mundt, an associate professor (førsteamanuensis) in the Department of Nordic Languages at Universitetet i Bergen, is presently preparing a computer-assisted study of the adverb in Old Norse.

frequencies of letters, word-structures, and related matters; to serve as an introductory guide to anybody wishing to carry out a similar study.

Apart from being written in a comparatively rare language, the book is not easily accessible, since the authors introduce a set of Icelandic terms for all the well-known features in computational linguistics. Even in the case of the young Icelandic student, who is supposed to read this book as an introduction, it is difficult to see just how long or to what extent he will benefit from these partly brand-new terms; as soon as he becomes interested in doing some sort of work in the field, he must cope with the internationally accepted set anyway. At least, in these circumstances, the book cannot very well be recommended as an 'introductory guide' to anybody outside Iceland.

The novel *Hreiðrið* (The Nest) by Ólafur Jóhann Sigurðsson (Reykjavík, 1972, 53 226 running words) was punched on cards in 1973, the machine being an IBM 26, which even then – as the authors admit with a smile – was already a kind of a museum piece.

During the following winter, in addition to the programming (in PL/1), they did some test runs, which convinced them that the IBM 1620 to which they had access at that time was not powerful enough to cope with the task. When finally the computer was replaced by a better one, the IBM 370/135, and the material was transferred to magnetic tape, the productivity increased rapidly, so that by the end of 1976 virtually all the required listing and computing was done.

Unfortunately, the writing of the final report, accounting for achievements, obstacles, experiences, and some pieces of augmented linguistic insight, could not be started seriously until two or three years later, resulting in a publication which in more than one place contains details concerning techniques outdated in 1980. As a rule, these will do no harm, since the experienced reader will soon find out where to skip a passage or a page, but I wonder if this is true for the quite inexperienced reader as well. Chapter 5, for instance, contains, in addition to the explanation of some basic terms such as files, primary and secondary storage, and so on, an account for the way the bytes of EBCDIC correspond to hole patterns on punch cards. When the newcomer, after careful reading of the pages 40–45, eventually understands how the research team managed to make the EBCDIC applicable to modern Icelandic, he has of course learnt something, but it can hardly be called guidance for future projects, since the reader will see from a later passage, that this coding scheme is not in vogue any more: "with the introduction of the new ISO-standard for all the Icelandic characters, it can now effectively be considered as obsolete" (p. 141).

Chapter 6 contains some general statistics and samples from the different kinds of listing carried out at various stages of the project. After a table displaying the frequency of the different types of graphic words and the total of letters, figures and logograms involved, we find samples from three word-frequency lists published in 1975, with types (that is, different running words) arranged in alphabetical order, reverse alphabetical order, and order of frequency. Listed separately we find the groups of words not made up of letters only (hybrid words), and finally the list of juncture graphemes, this last being published here for the first time.

Chapter 7 contains the bulk of quantitative findings from numerous tests run to serve the main task, the concordance. Some additional tests show how the investigation of letter-combinations can lead to the perception of structural patterns, which for obvious reasons hardly could have been made the object of thorough research in former times. For people not especially interested in the novel *Hreiðrið* Chapter 7 is the most important one. Most readers, I guess, will therefore be glad to see, that the better part of the English summary covers this section of the book.

As we know from the introduction of the book, Chapter 7 is almost entirely the work of Sven Þ. Sigurðsson, a mathematician by profession. Both his longlasting collaboration with Baldur Jónsson, an experienced philologist, and his own personal competence left their mark on what is said in this chapter, the arguments being as sensible and clear-cut as possible, even in the cases where more sophisticated statistical methods (such as cluster analysis) are employed. All of the eight tables presented in Chapter 7 are followed by some sort of

useful comment, not as exhaustively assessing all the features involved, but as giving valuable hints to the role that the respective kind of information may play, or already has played, in more recent research. Thus Table 1 shows the frequency and rank of the 100 most common words in *Hreiðrið*, as compared with the corresponding values for the only previous systematic word-count of modern Icelandic, i.e. the corpus of Ársæll Sigurðsson from about 1940 (cf. p. 20). Apart from explaining the figures in the table, the comment touches upon another computer-made word-count, which to some extent can be compared with the present one, Michael Bell's concordance to the Old Norse Egils saga Skalla-Grímssonar, and it touches upon Zipf's Law, which gets further support from almost all the words in the list under discussion, but not from the most frequent one.

The tables that follow supply the reader with figures for, among other things, the relation between word-length and frequency, the relative and the cumulative frequencies of the single letters, figures for the most common consonant–vowel-structures (on the basis of all words of length up to and including 15 graphs) and finally, in Table 8 are given the letter-transition frequencies, including frequencies of letters immediately preceding or following a blank. As an example of how more specialized work can be done on the basis of information included in Table 8, the distributive correlation between any two letters of the alphabet (and the blank) is calculated. The correlation coefficients, which are based on frequencies of letter pairs in the set of different words (types) are not shown as such in the report, but are summarized in the form of a dendrogram (p. 119). Another calculation based on the letter-transition frequencies, which will be highly interesting for at least some of the readers, are the one- and two-letter entropy values, shown at the bottom of the table itself (p. 114).

Chapter 8 contains information about the concordance, which obviously still exists in three copies only, printed in 1976 and equipped with a preface by Baldur Jónsson in 1978. Since the concordance does not come forth as a companion volume, we can dispense with a lengthy description of it. Nevertheless potential users may want to know that, from the samples as well as from the comment it is clear, the concordance has the form of a KWIC-index, with all the letters appearing as capitals, although we know from an earlier section of the report (p. 33) that the text was punched in a way which made it possible to keep apart capitals and small letters, provided that the hardware could take care of all the additional signs.

As a whole, the report suffers greatly from the gap between the levels on which the different sections are written: some seem useful merely for readers who do not have the slightest idea of computer-aided research, while other sections are at so high a level that only advanced readers can be expected to grasp the purpose of the certain routines. Despite the disparity in scientific level, however, the report is worth reading, since it gives a fairly objective picture of what has been going on in the field in this country during the last decennium. No parsing and no lemmatizing or anything of the kind seems to be done so far, and the indices do not aim at more than being lists of graphic words, without any attempt to distinguish homographs. Still, remembering that the three members of the team started the whole thing less than ten years ago, with less than the best technical device at their disposal, we have to take the book for what it is: a report accounting for the first 6–7 burdensome years with computers employed in language research in Iceland.

The NIV Complete Concordance, Edward W. Goodrick and John R. Kohlenberger III, Zondervan, Grand Rapids: 1981; xii +1044 pp.

By H. Van Dyke Parunak

This concordance is the first of a series planned by the Zondervan Publishing House for its *New International Version* of the Bible. It is the simplest

H. Van Dyke Parunak, an assistant professor in Near Eastern Studies at the University of Michigan, directs the Michigan Project for Computer-Assisted Biblical Studies, an archive of machine-readable biblical texts, and is interested in data base designs for storing and retrieving literary and linguistic data.

sort of concordance: a list of contexts and references, for a selection of the words occurring in the text, sorted in their inflected form.

Not every word of the NIV is concordanced. The word 'complete' in the title means only that every occurrence is listed of those words which are included. Such a strategy is an understandable way to reduce the bulk of a concordance, if function words such as 'and' and forms of 'to be' are eliminated. But the rationale for including and excluding words in this concordance is not so straightforward. The omitted forms are listed in an appendix. There we learn that 'DAUGHTER-IN-LAW' and 'SON-IN-LAW', important terms for the study of Hebrew social structures, are not concordanced, though 'BROTHER-IN-LAW', 'SISTER-IN-LAW', 'FATHER-IN-LAW', and 'MOTHER-IN-LAW' are. On the other hand, 28 pages are devoted to a listing of 'LORD' and 'LORDS', terms which are so ubiquitous as to be useless to a reader seeking a forgotten reference. 'SON' is concordanced, but the rarer 'DAUGHTER' is not. If either word is to be omitted, the less common one should be retained. The work shuns many adverbs, concordancing 'TENDER' but not 'TENDERLY', 'TERRIBLE' but not 'TERRIBLY', 'DECEIT' and 'DECEITFUL' but not 'DECEITFULLY', and so on. Such adverbs have considerable relevance to the semantic study of their base forms, and hardly warrant omission as being of 'very limited value in a concordance', as the preface claims (p. vi).

The words are sorted in their inflected forms. Thus one finds 'DRY', 'DRIED', 'DRIES', and 'DRYING' spread over two non-contiguous pages. Nouns are separated from their plurals, so that 'HEAD' and 'HEADS' are distinct (and non-contiguous) articles. Consistently with this policy, verbs and nouns that are spelled the same, such as 'FIGHT' or 'VISIT', fall together in a single listing. The economy of doing without analysis is thus purchased at the expense of considerable inconvenience and linguistic anomaly for the user, who must chase a single verb over several different entries while sorting through a single entry for different parts of speech.

To alleviate the problems caused by an unlemmatized text, the editors have listed a base form in parentheses after the heading of each inflected entry, and have listed after the heading of each base form all the inflections derived from it. This cross listing includes compounds and other derivations as well as inflections of verbs. Thus at the entry for 'FEED' we find '(FED FEEDING FEEDS OVER-FED PASTURE-FED STALL-FED WELL-FED)'. The entry for each of these in turn includes a reference back to 'FEED'. Negatives are not cross-referenced. 'BLAME' and 'BLAMELESS' are not marked as related, nor are 'END' and 'ENDLESS'. The preface duly notes this policy decision, but gives no rationale for it.

It is regrettable that some of the expense saved by the availability of the computer was not spent on a simple preliminary lemmatization of the text, which would greatly enhance the value of the present volume at comparatively low cost. The cross-references provided among articles would furnish the basic dictionary for an automatic lemmatization program. The further discrimination of homographs is more difficult, but was in fact partially implemented in weeding out omitted words. Thus 'WELL' is cataloged when it is a noun or adjective, but not as an adverb; 'OWN' as a verb, but not as an adjective, appears in the concordance; the noun 'LEAVES', but not the verb, is included. Finally, the publishers will need to do a thorough lemmatization sooner or later for some of the other volumes projected for this series.

It is the context that provides clues to the usage of the word being concordanced. Thus, in general, the longer the context presented with each word, the more useful a concordance will be. The preface notes that "the computer automatically counted the width of each letter so that the contexts would not exceed the line length" (p. vii). Apparently, this computation was applied after the contexts had been selected by some other means, for many of the lines are not filled to capacity. Most of the short lines end where they do because punctuation occurs. This is a reasonable criterion for stopping a context. But other contexts stop short where no such reason is apparent. This is particularly harmful when the target word is a verb governing a preposition, and the preposition is omitted.

For instance, Neh 4:4 reads, "Turn their insults back on their own heads." A case analysis of 'to

turn' would profit greatly by the knowledge that the verb governs a prepositional phrase introduced by 'on'. In fact, in the article for 'INSULTS', all but the word 'heads' is cited (p. 452). The inclusion of 'heads' would have overflowed the line. The entry for 'TURN', though, stops with the word 'back' (p. 963). The rest of the sentence, including the crucial preposition, is missing, even though there is room for all but the last two words.

Similarly, Matt 5:25, describing a lawsuit, reads in part, "or he may hand you over to the judge, and the judge may hand you over to the officer." The article for 'HAND' contains two lines for this verse. One of them reads, "or he may *h[and]* you over to the judge". The other is, "and the judge may *h[and]* you". Semantically, 'to hand' and 'to hand over' are quite distinct. The omission of 'over' in the second case is misleading, suggesting that the two uses of 'hand' in Matt 5:25 are distinct. This omission is also unnecessary, since there is more than enough room left on the line to include 'over', or even 'over to'.

The work would be much more useful, and not much more expensive, if the contexts were presented in KWIC format, sorted in order of following context rather than in order of occurrence. If the traditional order were deemed necessary, each article could include a separate list of references without contexts. This would not add greatly to the size of the work. Even if the entries were given only in citation order, the KWIC format would still be helpful.

The NIV Complete Concordance is a traditional tool of considerable usefulness whose prompt production was made possible by the computer. Hopefully, those who design its successors in the series proposed by Zondervan will produce state-of-the-art computer concordancing.

Maschinelle Sprachanalyse, Peter Eisenberg, ed., Beiträge zur automatischen Sprachbearbeitung I, Walter de Gruyter, Berlin, New York: 1976.
Semantik und künstliche Intelligenz, Peter Eisenberg, ed., Beiträge zur maschinellen Sprachbearbeitung II, Walter de Gruyter, Berlin, New York: 1977.

By Gerd Willée

The aims of both volumes – as they are expressed by the editor and by the publisher – are:
- to give a survey on the latest developments in the fields of automatic language analysis, computer semantics, and artificial intelligence,
- to encourage linguists to use the computer as a tool for their research, and to use the results of computational linguistics,
- to make available papers and reports on these topics which would not be otherwise accessible,
- to present an anthology which can be used as textbook in teaching at university level.

Before going into details about the books it must be stated that these goals have been attained, or, as for the second goal, should be attainable. As there are no publications in West Germany on the same subjects which could be compared with Eisenberg's books and which could be used in a similar way, these books really fill a gap.

A condensed introduction by the author is followed by the selected articles, which discuss special problems or give surveys over more general matters.

Each book finishes with a bibliography related to the articles and enlarged by other relevant titles, followed by a detailed subject index. From the didactic point of view, this structure, together with the selection of the articles, makes Eisenberg's books very handy and highly suitable for university courses.

Maschinelle Sprachanalyse

This volume is a good introduction to computational linguistics, as the eight articles contained in it give a non-trivial overview of computer use in linguistics, i.e. emphasis has not been on text processing, like establishing concordances or indices, but on text analysis, on parsers, application of grammars and the like.

The anthology opens with a contribution of Hans Karlgren and Benny Brodda on 'The computer as a tool for resolving problems which cannot be formalized', in which the interdependency between

Gerd Willée works in the fields of computational linguistics at the University of Bonn.

traditional linguists and computational linguists is discussed, as well as the future role of the computer in linguistic research, against the background of the experience with the computer from the enthusiastic hopes in the beginning of computational linguistics. This article has an useful side effect, as the reader gets an idea of the computational linguistics being done in Scandinavia, which, because of the language barrier, one is normally not able to follow.

The next article is a translation of Joyce Friedman's 'A computer system for transformational grammar' followed by Istvan Bátori's 'Teleology of types of grammar: Generative grammar and analytical grammar', Eberhard Pause's 'Tests of adequacy and syntax analysis', and the translation of William Wood's 'Transition network grammars for natural language analysis'.

Wolfgang Klein's 'Computational analysis of linguistic change: A descriptive method demonstrated on the high German sound shift' gives an example for how computers can be used in historical linguistics. Klein points out that applications in this field can be made more easily than in other fields of linguistics, as historical linguistics mainly deals with the phonetics, phonology, morphology, surface structure syntax, and lexicology of the stages of the language under investigation. Moreover the databases are comparatively limited. Two reasons he gives for there being only a few applications limited to the production of concordances, wordlists, and the like are the lack of interest in such topics among the computational linguists and the lack of suitable methods, in other words, a theoretical concept which adequately describes the variations and changes of languages. He then describes the methods by means of which he made an analysis of the sound shifts in Early Modern German using computer programs, showing at the end of his essay that using a computer can be worthwhile for some purposes in historical linguistics.

Susumu Kuno's 'Computer analysis of natural languages' gives a survey on the methods used for parsing by different research groups.

Robert F. Simmons' 'Natural language question answering systems', like Kuno's article a translation, can be looked at as a link to Eisenberg's second volume.

Semantik und künstliche Intelligenz

As there are lots of unpublished papers in the fields of AI, five out of eight articles in this volume are here published for the first time, giving the reader an idea of what is going on in the fields of simulating language behavior and understanding language.

The book starts with Marvin L. Minsky's 'Matter, minds, and models' (1968) as a basic approach to the principles of constructing intelligent machines.

Eugene Charniak's 'Reference and question answering in simple narrations' shows some of the problems one faces when leaving well-defined miniworlds and using narrations for question answering systems, which are easily understood by 8-year-old children, but which, however, turn out to have a quite complex structure.

Bruce Fraser presents 'Some pessimistic views on improving man–machine communication' in which he corrects his more optimistic outlook from 1965 on possible future computer systems, with which a non-programmer easily would be able to communicate, showing the problems of taking into account pragmatic aspects and similar matters, when understanding language.

Peter Hellwig's 'A computer model for inferences in natural language' explains the possibilities and principles of an efficient deduction procedure, which operates with natural language data.

David G. Hays' essay 'Cognitive networks: Forms and processes' describes attributes of a cognitive network with cognitive, sensomotorical, and language faculties. In 'Computers, primitive actions, and linguistic theories' Roger G. Schank gives a description of some features of MARGIE, showing how semantic representations can be stored in the form of elementary concepts, and pointing out that all concepts of action can be covered by twelve primitive actions.

Terry Winograd's 'A procedural model of language understanding' is a translation of an essay from 1973 which describes SHRDLU and its block world together with some of its main concepts and programs working with them.

The volume closes with a revised version of Yorick Wilks' 'Natural language understanding

systems within the A.I. paradigm: A survey and some comparisons' (1974). This essay discusses some aspects of the developments and tendencies in AI, making – as a sort of an additional chapter to the papers presented by Winograd, Schank, and Charniak – some critical notes on their models, in comparison with those of Colby, Simmons, and Wilks.

It must be pointed out that Eisenberg's books, based on a good concept, are useful not only to the German reader.

A Guide to Computer Applications in the Humanities, Susan Hockey, The Johns Hopkins University Press, Baltimore, MD and London: 1980.

By Raoul N. Smith

On the whole, this lucidly written introduction to computers in literature and linguistics is a good survey of various topics of interest to a great many people. It can serve as a textbook for both beginning students and researchers in the field. The title is misleading, however, since it does not deal with at least two major representatives of the humanities, namely art and music, and describes only briefly projects in archaeology, bibliography, the law and history. The majority of the studies described, in literature and a few areas in linguistics, all deal with surface text processing.

Based on Hockey's lectures at Oxford, the book was written primarily for a British audience. She writes, for example, "Instructions are presented to the computer in the form of a computer *program* (always spelt 'program')." In the somewhat dated Chapter 2, 'Input and output', she spends eight pages (22–30) on keypunch machines, punch cards and paper tape and then briefly mentions VDUs, the standard data entry device today. Somewhat anachronistically, she states, "Some line printers have upper and lower case letters" (p. 27), refers to multi-ply paper for copies, and makes no reference to laser printers.

In Chapter 3, 'Word indexes, concordances and dictionaries', which deals mainly with concordances, she does a good job of explaining and illustrating them, especially those of bilingual texts. The illustrations, however, could be compressed. For example, the left page of the two-page Figure 3.6 has no examples of what is the topic of the figure, and since Figure 3.7, also two pages long, derives from 3.6, no examples appear on the first page of that pair either. The discussion of concordances of the non-alphabetic symbols of Minoan Linear B (one paragraph on p. 69) is treated as sort of an aside, when it would have been a good opportunity to discuss the broader use of computers in the humanities in analyzing other kinds of symbol sets.

The fourth chapter, 'Vocabulary studies, collocations and dialectology', takes the next step from word indexes to treat word counts, although rather cursorily, given the amount of literature on this subject. The subsection on collocations is a welcome addition of a topic which is seldom covered in works on text processing. Similarly, the discussion of the computer in dialect research is a welcome addition to an often neglected area.

The chapter which should have followed this is Chapter 6, 'Stylistic analysis and authorship studies', since it touches on some of the statistical notions raised in Chapter 4. Again, Hockey does a good job of hitting the highlights of this voluminous, and at times difficult literature.

Chapter 5, which concerns 'Morphological and syntactic analysis, machine translation', begins with a discussion of lemmatization projects and then continues with the syntactic analyses of EYEBALL. This section adopts a critical stance rather than the reportive one taken in the rest of the book. Again there are lots of sample printouts but these are useful for the beginner. The transition into the section on machine translation, while it is abrupt, nevertheless makes an important point that bears repeating: machine translation will not be possible until very large dictionaries of collocations are available. Two annoying aspects of the book are exemplified in this chapter: the 'references' at the end of each chapter are often not referred to

Raoul Smith, formerly a professor of linguistics at Northwestern University, is a senior member of the Technical Staff at GTE Laboratories, where he is researching the linguistic aspects of human–computer communication.

in the text, and references are sometimes made in the text to works, such as that by J.B. Lovins (p. 103), which are not listed in the references.

Chapters 7 and 9, on textual criticism and on indexing, cataloguing, and information retrieval, should be sequenced that way, with Chapter 8 (on sound patterns) placed after. In Chapter 9, there is a description of how the index to the book was constructed, namely, the text was read and a list of items with their page numbers compiled. These were then entered into the computer and sorted. This is an unfortunate example since there are much easier and imaginative methods by which the computer could have been used in creating the index.

The last chapter, 'How to start a project', is a collection of odds and ends on where and how to get information and help about setting up a project.

As can be judged from the foregoing comments, *A Guide to Computer Applications in the Humanities* is an introductory survey of projects involving computers and literary texts which will be of benefit to the neophyte. A set of in-depth case studies, as a companion volume, would make a useful accompaniment to it and form a powerful resource for introductory courses in this field.

Index-concordance d'Emile ou de l'éducation, tomes I et II, Etienne Brunet, Slatkine, Genève et Champion, Paris: 1980; LX +585 pages et XXI +727 pages.

Par Serge Lusignan

Voilà un magnifique travail que nous livre Etienne Brunet dans deux forts volumes publiés chez Slatkine et Champion. Avant d'étayer mon jugement, j'aimerais saluer ici la contribution générale de Brunet à la constitution de cette collection qui publie des ouvrages méthodologiques et des résultats d'analyse dans le domaine du traitement des textes par ordinateur. Cette série est en train de devenir un des points de référence pour ceux qui s'intéressent à la recherche de nos collègues français dans ces domaines. Mais revenons à l'index-concordance du grand traité de Jean-Jacques Rousseau sur l'éducation.

Le tome I s'ouvre par deux longues dissertations de Michel Launay, l'une intitulée 'Emile au 20e siècle ou: Rousseau, Freinet, Gransci', l'autre 'Gransci et le rousseauisme'. Suivent une introduction de Brunet et l'index lui-même qui forme le corps du volume. Le livre se termine par une liste des mots de fréquence élevée, le tableau de la distribution des fréquences et deux courtes études statistiques très intéressantes: l'une sur la distribution du vocabulaire à travers les cinq livres de l'*Emile*, l'autre sur le rythme du texte saisi à travers la distribution de la ponctuation. Le tome II quant à lui contient essentiellement une concordance sélective de l'*Emile* basée sur une étude du vocabulaire spécifique du traité.

Le travail d'Etienne Brunet se recommande autant aux spécialistes de Rousseau, cela va de soi, qu'à ceux qui s'intéressent à la méthodologie des concordances et des relevés lexicaux en général. Il présente un modèle, qu'on peut discuter et c'est ce qui en fait sa richesse, mais qui sort ce genre d'ouvrage d'une stérilité intellectuelle qui en est trop souvent la principale caractéristique.

Au-delà de leur usage connu et utile pour le dépistage de l'information, les index-concordances et relevés quantitatifs publiés suggèrent régulièrement la pertinence du modèle statistique pour les études littéraires. Il est devenu courant dans nos discussions sur le style d'un auteur d'intégrer les notions de fréquence des vocables, de richesse du vocabulaire, de distribution des fréquences. Pourtant nous sommes encore bien incertains quant à la détermination et à la délimitation des paramètres quantitatifs d'un texte. Nous ne savons pas avec précision comment devraient se construire les séries statistiques pour l'étude des textes. A ce niveau, nous en sommes aux balbutiements si on compare aux sciences économiques par exemple.

Le livre de Brunet prend à ce sujet un certain nombre d'options qui imprègnent toute sa démarche et tout d'abord sur les unités de base qui sont comptabilisées pour fonder les études statistiques. Brunet prend le parti de ne pas lemmatiser alléguant nos incertitudes quant à la langue du

Serge Lusignan est professeur d'études médiévales à l'Université de Montréal.

XVIIIe siècle et la perte d'information qu'entraîne la réduction des formes à un vocable. Ces arguments sont recevables à défaut d'être très contraignants. S'il est vrai que la lemmatisation opère une certaine réduction, l'information élaguée peut néanmoins être conservée par une présentation hiérarchique (ventilation des vocables selon leurs différentes formes) des index ou des comptes de vocabulaire. La raison la plus valable de ne pas lemmatiser dans le cas d'un ouvrage comme celui de Brunet me semble être le peu d'inconvénients qu'il y a à faire l'économie de cette très lourde entreprise: un index ou une concordance non lemmatisés sont relativement facile d'emploi. Où la perte est plus grave, c'est au niveau de l'information linguistique: les index ou concordances non lemmatisés sont nettement moins riches. Mais cela nous ramène à l'autre argument de Brunet pour ne pas lemmatiser à savoir qu'il ne voulait pas faire oeuvre de pionnier dans la description du français du XVIIIe siècle, laissant ce projet aux chercheurs de l'I.L.F. (tome I, p.LIII).

Cette suspension de jugement mène loin lorsqu'on l'applique à un texte français du XVIIIe siècle. L'orthographe présente une certaine instabilité surtout au niveau de l'accentuation. Par exemple 'élève' peut s'écrire: éléve, élêve, éleve ou eleve dans Rousseau. Brunet décide à juste titre de ne toucher à rien, même s'il en résulte une surabondance de formes différentes, en raison de sa volonté de ne pas interpréter le texte au plan linguistique.

Ayant déterminé les éléments de base qui sont comptabilisés dans son index-concordance, Brunet organise les données qu'il nous présente selon deux séries statistiques. La première série est construite à partir du décompte séparé des fréquences des formes pour chacun des cinq livres de l'*Emile*. Autrement dit, Brunet postule que toutes les mesures de ce texte doivent tenir compte de sa division en livre. Cette série statistique sert de structure à l'information qui contient le premier tome de son travail.

La principale conséquence de cette décision se lit dans la présentation de l'index. Chaque forme est accompagnée de ses références présentées en cinq colonnes, chacune correspondant à un livre. Le résultat est intéressant car il donne à l'index une double fonction: informer sur les lieux d'occurrences de chaque forme et montrer la variation d'emploi de la forme à travers les cinq livres. Il indique aussi la fréquence de la forme dans chaque livre, lorsqu'elle est égale ou supérieure à cinq, et sa fréquence totale. Cette façon d'organiser l'index est fort utile tout autant que claire.

L'index principal recense toutes les formes dont la fréquence ne dépasse pas 100 dans au moins un livre. S'ajoute à cela un index complémentaire pour 15 formes plus fréquentes et riches sémantiquement, ce qui ne laisse finalement que 95 formes très fréquentes non indexées. Pour une oeuvre de la taille de l'*Emile* (258 501 mots) le compromis est raisonnable. L'index est complété par une liste de fréquences, globale et par livres, des formes les plus fréquentes et par une table de distribution des fréquences ventilée de la même façon, sauf pour les fréquences supérieures à 200. Pour ces dernières, le tableau ne s'occupe que des fréquences globales. C'est une décision que l'on comprend mal.

Etienne Brunet poursuit l'étude des données qu'il publie en complétant son tome I par une étude comparative de la richesse lexicale des cinq livres. Cette étude vient justifier a posteriori la ventilation des données qui précède. Elle démontre que de toute évidence il s'agit d'une structure quantitative réelle de l'*Emile* et fonde la validité des études futures qui s'appuieront sur ce paramètre. Le bien-fondé de cette structure est encore mis en évidence par une dernière étude sur la ponctuation. Elle montre comment la nature de celle-ci et le rythme qu'elle impose au texte varient selon les livres tout en étant assez homogène à l'intérieur de chacun, sauf peut-être le quatrième.

Au total on pourrait résumer l'apport du premier tome de travail de Brunet en ce qu'il fournit un outil valable de dépistage de l'information et une hypothèse bien étayée quant à la structure de la distribution quantitative du vocabulaire dans l'*Emile* de Jean-Jacques Rousseau. Sa contribution dépasse de beaucoup celle des fabriquants d'index et c'est par là qu'elle a valeur d'exemple.

Après avoir étudié la structure interne de l'oeuvre, Brunet essaie dans le tome II de caractériser l'*Emile* de Rousseau par rapport à un certain état de langue. Il s'agit en quelque sorte de le situer dans la *série* des oeuvres de son époque.

Ici Brunet s'aventure sur un terrain beaucoup plus fragile, principalement à cause du manque de données sur la question et par conséquent de notre ignorance. Son objectif est d'extraire le vocabulaire significatif de l'oeuvre en vue d'opérer les difficiles choix qui s'imposent à quiconque veut contenir dans des limites raisonnables la concordance d'un texte trop volumineux.

L'introduction du tome II présente la méthode et les résultats. Grâce au corpus du Trésor de la langue française (Nancy), Brunet a pu établir deux corpus qui lui servent de normes pour isoler le vocabulaire propre de l'*Emile*. Le premier (6 millions de mots) comprend des textes français de 1789 à 1815, le second (20 millions de mots) des textes de la première moitié du XIXe siècle. Bien sûr un corpus plus nettement du XVIIIe siècle aurait été préférable et Brunet en convient; mais c'est là l'état de la recherche dans ce domaine. Il a donc établi une liste de formes qui à l'aide d'un modèle probabiliste s'avéraient très nettement sur-représentées ou sous-représentées dans l'*Emile*. Une forme n'est retenue que si elle est hors norme par rapport aux deux corpus. Aux formes ainsi extraites, Brunet ajoute celles de même racine sémantique mais non détectées parce que de fréquence trop basse. C'est donc à partir de ces critères qu'il lui est possible de sélectionner 1571 formes significatives parmi les 14 559 que compte l'*Emile* et d'en fournir la concordance.

Ce deuxième tome du travail de Brunet est à la fois intéressant et stimulant, bien que sujet à davantage de discussion. C'est qu'il s'aventure sur un terrain beaucoup plus inconnu sur lequel on peut espérer que d'autres chercheurs n'hésiteront pas à le suivre. Car s'il est intéressant d'étudier quantitativement une oeuvre en ses parties ou relativement à d'autres oeuvres d'un même auteur, il faudra aussi avoir les moyens de la situer quant à ses paramètres quantitatifs dans un contexte plus général d'une époque, d'un genre littéraire.

Brunet ne pousse pas très loin la façon de définir une grande série dans laquelle insérer une oeuvre. Il prend une tranche chronologique. A la page VII il lui échappe la phrase un peu malheureuse de 'situer ce texte dans la langue et la littérature française'. Je crois qu'il faudra utiliser des notions beaucoup moins englobantes, beaucoup moins ambitieuses, mais beaucoup plus descriptives, pour définir la constitution de grandes séries quantitatives. Néanmoins, il nous indique le bon chemin en nous invitant à dépasser le contexte d'une oeuvre ou d'un auteur comme corpus témoin pour faire ressortir des spécificités. D'ailleurs le résultat de cette distillation des données est très fascinant et donne l'impression de pénétrer au coeur même des thèmes caractéristiques de l'*Emile*.

Il va de soi qu'après une telle sélection, le mot concordance qui apparaît dans le titre du travail n'est pas sans ambiguïté. On attend d'une concordance la recension de tous les mots ou au minimum de tous les mots 'significatifs' d'un texte. La concordance de l'*Emile* est loin de cet objectif. Elle serait plus justement appelée, matériaux pour l'étude du vocabulaire spécifique de l'*Emile*. Cela lèverait une ambiguïté et rendrait davantage justice au travail d'Etienne Brunet qui est rempli de raffinement et d'intelligence.