Computational and Quantitative Studies

M. A. K. Halliday and edited by Jonathan J. Webster
(City University of Hong Kong)

London: Continuum (The collected works of M. A. K. Halliday, volume 6),
2005, x+300 pp; hardbound, ISBN 0-8264-5872-6, $150.00

Reviewed by
Chris Mellish
University of Aberdeen

Michael Halliday is a linguist whose ideas have inspired computational linguists over a prolonged period. His earliest research work took place in China and with Chinese languages. In the late 1950s, as an assistant lecturer in Chinese at Cambridge, he worked in a project concerned with machine translation at the Cambridge Language Research Unit (directed by Margaret Masterman and which was later to host researchers such as Karen Spärck Jones and Yorick Wilks). The use of Halliday's Systemic Grammar was a major feature of Winograd's influential work (Winograd 1972) that did a great deal to put NLP on the map in the late 1970s. But Halliday's approach has probably had its most enduring influence on work in natural language generation, from Davey's pioneering work (Davey 1978) through the ISI PENMAN project (Mann 1983) to a number of recent NLG systems that are influenced particularly by it (Fawcett and Tucker 1990; Bateman 1997).

This book is Volume 6 of the collected works of Halliday, specifically devoted to his computational and quantitative work. His collected works are a retrospective of around 50 years of work. Halliday formally retired around 1987, but the 11 papers in this book date from 1956 to 2002, with at least seven coming from after 1990. The book is in three parts, covering his early years in MT, works on corpora and probabilistic grammar, and some papers about intelligent computing. My personal feeling is that the second of these is the part most of interest to computational linguists, the first being of restricted interest because so much has happened in MT in the last 50 years; and the last being perhaps overly oriented to the vision of Sugeno (1995) that intelligent computing could use natural language as a metalanguage (which Halliday does not specifically endorse or feel able to comment on in detail). The papers tend to be quite general, non-formal, overview/discursive articles, apart from Chapter 6, which presents a quantitative study with Z. James investigating the manifestations of polarity and primary tense in the English clause using the COBUILD corpus (Halliday and James 1993). A useful appendix to the book shows the grammar of the English clause that Halliday provided to the PENMAN project around 1978.

What has made Halliday's approach to grammatical description appealing to computational linguists are the importance it places on meaning/context, the concern with real language, and also the specific formalism he uses. As regards the last of these, the notation of Systemic Grammar has been formalized (Patten and Ritchie 1987) and theorems proved about its computational properties (Brew 1991). Halliday, however, believes that language is inherently fuzzy and so he probably resists this level of formality in the description. What computational linguists have not picked up on (perhaps because this was not so fashionable until more recently) is Halliday's

view that "a linguistic system is inherently probabilistic in nature." Halliday sees no distinction between corpus linguistics and theoretical linguistics, adapting the principle that "frequency in text instantiates probability in the system." But he criticizes early MT work that gets involved with "the counting of orthographic words in isolation" and believes that corpora should be as useful for grammarians as they have been for lexicologists. But "the grammar is much harder to get at" and "the main limitation on the use of corpuses [sic] for probabilistic grammar is the familiar catch that what is easy to recognize is usually too trivial to be worth recognizing." So the joint work with James referred to above has to use a number of complex heuristics (in a way that will be familiar to computational linguists) to extract the desired grammatical information from the corpus. Recent developments in wide-coverage parsing and the exploitation of huge corpora should mean that the boundaries are moving very fast in this area, if there are grammarians out there willing to have a go ...

There are a number of intriguing ideas and comments to be found in the papers. For instance, there is the hypothesis (with some empirical support) that binary choice systems in languages tend to have either equal probabilities for the two options ("equiprobable") or with probabilities around 0.1 and 0.9 ("skew"). Interestingly, the latter case corresponds to where the information-theoretic entropy is about 0.5 (which is also about the figure for the entropy for English in terms of *characters*, as calculated by Shannon and Weaver). Also there are some interesting speculations about how language changes that manifest themselves initially through probabilities might eventually lead to changes in the underlying systems themselves.

Although I don't think that computational linguists will find any ideas in this book that will directly inspire computational implementation, I found it an interesting read that gave me a number of insights into the history of linguistics and into how Halliday's thinking relates to computational issues. Halliday's writing is erudite and clear, but it is also quite dense and uses terminology that has to be mastered. The examples and grammar fragments help a lot to make the ideas precise, but I'd have liked there to be more of these.

References

Bateman, John. 1997. Enabling technology for multilingual natural language generation: The KPML development environment. *Natural Language Engineering*, 3(1):15–55.

Brew, Chris. 1991. Systemic classification and its efficiency. *Computational Linguistics*, 17(4):375–408.

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

Fawcett, Robin and Gordon Tucker. 1990. Demonstration of GENESYS: A very large, semantically based systemic functional generator. In *Proceedings of the 13th International Conference on Computational Linguistics*, pages 47–49, Helsinki.

Halliday, Michael and Zoe James. 1993. A quantitative study of polarity and primary tense in the English finite clause. In John Sinclair, Michael Hoey, and Gwyneth Fox, editors, *Techniques of Description: Spoken and Written Discourse*. Routledge.

Mann, William. 1983. An overview of the PENMAN text generation system. In *Proceedings of the National Conference on Artificial Intelligence*, pages 261–265. American Association for Artificial Intelligence, August.

Patten, Terry and Graeme Ritchie. 1987. A formal model of systemic grammar. In Gerard Kempen, editor, *Natural Language Generation: New Results in Artificial Intelligence, Psychology and Linguistics*. Martinus Nijhoff.

Sugeno, Michio. 1995. Intelligent fuzzy computing. In *Proceedings of PACLING II (Second Conference of the Pacific Association of Computational Linguistics)*, Brisbane.

Winograd, Terry. 1972. *Understanding Natural Language*. Edinburgh University Press, Edinburgh.

Chris Mellish has been working in Computational Linguistics since his Ph.D. research at the University of Edinburgh, which was published in 1981. After a period at the University of Sussex, he returned to Edinburgh as a lecturer and in Edinburgh was appointed to a chair in Natural Language Processing in 2001. Mellish is now at the Department of Computing Science, University of Aberdeen, King's College, Aberdeen AB24 3UE, UK. His e-mail address is cmellish@csd.abdn.ac.uk.