

promoting access to White Rose research papers



Universities of Leeds, Sheffield and York
<http://eprints.whiterose.ac.uk/>

This is an author produced version of a paper published in Journal of Information Science

White Rose Research Online URL for this paper:
<http://eprints.whiterose.ac.uk/5417/>

Published paper

Willett, P. and Robertson, S. (2007) *In Memoriam: Karen Sparck Jones*. Journal of Information Science 33(5) pp.637-639

IN MEMORIAM: KAREN SPÄRCK JONES

Peter Willett and Stephen Robertson

Karen Spärck Jones, one of the pioneers of information retrieval (IR), died on 4th April 2007 at the age of 71. Her career history can be stated simply. She was born in Huddersfield in 1935 and brought up there before going to Cambridge University in 1953 to read history. With the exception of a brief period as a teacher after graduating, she spent her entire life at the University, initially as a PhD student and then research worker in the famous Cambridge Language Research Unit under the supervision of Margaret Masterman. She was awarded her doctorate in 1964, her thesis, *Synonymy and Semantic Classification*, having the distinction of being republished in full by Edinburgh University Press over twenty years later, by which time the importance of her doctoral research had become evident to the wider community (Spärck Jones, 1986). Following her time with the Unit, she was successively a Research Fellow of Newnham College, Royal Society Research Fellow, Senior Research Associate, GEC Fellow and then Assistant Director of Research. She only achieved a permanent position in the University in 1994 when she was made Reader in Computers and Information, and was finally awarded her chair in 1999, just three years before her retirement. She was a Fellow of the American Association of Artificial Intelligence (AAAI) and of the European Coordinating Committee for Artificial Intelligence, and the President of the Association for Computational Linguistics (ACL) in 1994. Her major awards included the Gerard Salton Award of the Association for Computing Machinery (ACM) SIGIR in 1988, the ACL Lifetime Achievement Award in 2004, and the British Computer Society Lovelace Medal and the ACM/AAAI Allen Newell Award in 2007. She was also the recipient of a substantial festschrift, containing contributions from around the world from those who had worked with her throughout her career (Tait, 2005).

Karen's PhD had involved the construction of a thesaurus using text distribution data and statistical classification methods, and this led directly to her pioneering studies of keyword classifications. The work involved grouping words on the basis of co-occurrence data, with the aim of enhancing, principally, the recall of IR systems by allowing matches on clusters of words, rather than the just the individual words in a query or document (Spärck Jones, 1971). Although this work was ultimately unsuccessful, it involved the development of a systematic approach to the testing and evaluation of IR systems, with rigorous experimental design being used to control the many variables that can affect retrieval effectiveness. This detailed approach, which is clearly demonstrated in the two major British Library projects that she

undertook in the Seventies (Spärck Jones and Bates, 1977; Spärck Jones and Webster, 1980), characterised all of Karen's work and set a standard for researchers in the field that they needed to aspire to if their results were to be taken seriously. Her interest in experimental design was further demonstrated by her work on the design of document test-collections for IR experiments (e.g., Spärck Jones and van Rijsbergen, 1976), studies that did not have any immediate outcome but that provided clear guidelines for the TREC (Text Retrieval Conference) datasets that started to appear in the early Nineties. This interest in methodology led to her editing, and contributing two of the chapters in, *Information Retrieval Experiment*, one of the key publications in the emergence of IR as a discipline (Spärck Jones, 1981).

The work on keyword classification had suggested that weighting could be at least as effective as classification. This observation led to Karen's extensive studies of index-term weighting schemes, these resulting in two of the most important components of modern IR systems: IDF weighting (Spärck Jones, 1972) and then relevance weighting (Robertson and Spärck Jones, 1976), with the success of the latter providing some of the best evidence for the probabilistic models of retrieval that were becoming established at that time. The arrival of TREC in 1992 saw the realisation of her thoughts over the years on the design of IR test collections and IR experiments, and her extensive involvement in this continuing series of conferences included both significant contributions to the OKAPI BM25 weighting scheme (Spärck Jones, *et al.*, 2000), now widely recognised as the state-of-the-art in index-term weighting, and extensive involvement in the planning and implementation of the TREC programme (Spärck Jones, 1995). The Nineties also saw her playing a leading role in the development of techniques for multimedia retrieval, focussing on speech and video retrieval (Spärck Jones *et al.*, 1996; Tuerk *et al.*, 2001). She was also the editor of *Readings in Information Retrieval* (Spärck Jones and Willett, 1997), containing many of the key papers in the development of IR and providing a textbook function for the field. It says much for Karen's focus on the "tough" topics of the moment that much of her later research focused on question answering, automatic summarisation and multimedia retrieval, three of the eleven research challenges identified as facing the IR community over the next few years at a 2002 international workshop (Allan and Croft, 2002); a fourth such challenge was post-TREC datasets, a further area in which she would have undoubtedly have contributed.

This is a record of achievement of which anybody could justifiably be proud. However, Karen had built for herself a reputation at least as significant in a different field, that of natural language processing. In particular, she carried out important work on natural language access to databases, belief revision, question answering systems and automatic summarisation (see, e.g., Cawsey *et al.*, 1992; Copestake and Spärck Jones, 1990; Spärck

Jones and Endres-Niggemeyer, 1995; Spärck Jones and Tait, 1984). Many of these studies, as well as the writing of several books (Grosz *et al.*, 1986; Spärck Jones and Kay, 1973; Spärck Jones and Wilks, 1983), were carried out whilst she was also heavily engaged in IR research. It is thus appropriate that her final publications included one on TREC (Spärck Jones, 2006) and one on summarisation (Spärck Jones, 2007), the two main research areas in the final part of her career.

Karen was highly productive, her 229 publications (which are listed at <http://www.cl.cam.ac.uk/~ksj21/ksjdigipapers/ksjbib3.html>) including ten books and 69 journal articles. Her publications are heavily cited in the literature: there are currently over 1500 citations in *Web of Science* and over 4500 citations in Google Scholar, with Robertson and Spärck Jones (1976) and Spärck Jones (1972) being the most cited items in both databases. She published widely, but no less than eleven of her papers were published in the *Journal of Documentation*, with two of them subsequently being reprinted in the series of papers “60 years of the best in information research”. She was a strong supporter of the *Journal*, not just submitting papers but also being on the editorial board from 1975-1985 and being the Managing Editor from 1977-1980, thus helping other authors to meet the high standards that she set herself.

Karen brought an extraordinary and infectious energy and enthusiasm to her life. With her companion, and husband of 45 years, Roger Needham (who died in 2003), there was indeed a true companionship. They had no children, but at the beginning of their married life they built with their own hands a house just outside Cambridge, while working on their respective theses and living in a caravan on the site. Later they acquired, restored and sailed boats on the east coast, in particular an Itchen Ferry Cutter built in 1872. They travelled widely, and Karen always took great interest in the history, art and architecture of any place she happened to visit. One might run into her in a strange city sketching architectural details, and she would always enthusiastically suggest places to visit if one were about to travel to some destination with which she was familiar. Her approach to her working life was equally vigorous: identifying new avenues for exploration, probing untested assumptions, suggesting additional experiments, and highlighting analogous or related research. Collaborating with her might sometimes be tiring but it was always inspiring.

Karen’s work has profoundly influenced the design of IR and NLP systems. She has thus benefited not just the members of these academic communities but also the millions of people who interact with text-based information systems each day via the Web.

REFERENCES

- Allan, J. and Croft, W.B. (editors) (2002), *Challenges in Information Retrieval and Language Modeling. Report of a Workshop held at the Center for Intelligent Information Retrieval, University of Massachusetts Amherst, September 2002.* At URL <http://www.acm.org/sigs/sigir/forum/S2003/ir-challenges2.pdf>
- Cawsey, A., Galliers, J., Reece, S. and Spärck Jones, K. (1992) "Automating the librarian: belief revision as a base for system action and communication with the user", *Computer Journal*, Vol. 35, pp. 221-232.
- Copestake, A. and Spärck Jones, K. (1990) "Natural language interfaces to databases", *Knowledge Engineering Review*, Vol. 5, pp. 225-249.
- Grosz, B., Spärck Jones, K. and Webber, B. (1986), *Readings in Natural Language Processing*, Morgan Kaufman, Los Altos CA.
- Robertson, S.E. and Spärck Jones, K. (1976) "Relevance weighting of search terms", *Journal of the American Society for Information Science*, Vol. 27, pp. 129-146.
- Spärck Jones, K. (1971), *Automatic Keyword Classification for Information Retrieval*, Butterworths, London
- Spärck Jones, K. (1972) "A statistical interpretation of term specificity and its application in retrieval", *Journal of Documentation*, Vol. 28, pp. 11-21.
- Spärck Jones, K. (1981), *Information Retrieval Experiment*, Butterworths, London, 1981.
- Spärck Jones, K. (1986), *Synonymy and Semantic Classification*, Edinburgh University Press, Edinburgh.
- Spärck Jones, K. (1995) "Reflections on TREC", *Information Processing and Management*, Vol. 31, pp. 291-314.
- Spärck Jones, K. (2006) "What's the value of TREC – is there a gap to jump or a chasm to bridge?", *SIGIR Forum*, Vol. 40(1), pp. 10-20.
- Spärck Jones, K. (2007), "Automatic summarising: the state of the art. *Information Processing and Management*, in the press.
- Spärck Jones, K. and Bates, R.G. (1977), *Research on Automatic Indexing 1974-1976*. British Library Research and Development Report no. 5465.
- Spärck Jones, K. and Endres-Niggemeyer, B. (1995) "Introduction: automatic summarising", *Information Processing and Management*, Vol. 31, pp. 625-630.
- Spärck Jones, K., Jones, G.J.F., Foote, J.T. and Young, S.J. (1996) "Experiments in spoken document retrieval", *Information Processing and Management*, Vol. 32, pp. 399-417.
- Spärck Jones, K. and Kay, M. (1973), *Linguistics and Information Science*, Academic Press, New York.
- Spärck Jones, K. and Tait, J.I. (1984) "Automatic search term variant generation", *Journal of Documentation*, Vol. 40, pp. 50-66.
- Spärck Jones, K. and van Rijsbergen, C.J. (1976) "Information retrieval test collections", *Journal of Documentation*, Vol. 32, pp. 59-75.
- Spärck Jones, K., Walker, S. and Robertson, S.E. (2000) "A probabilistic model of retrieval: development and comparative experiments", *Information Processing and Management*, Vol. 36, pp. 779-840.
- Spärck Jones, K. and Webster, C.A. (1980), *Research on Relevance Weighting 1976-1979*. British Library Research and Development Report no. 5553.
- Spärck Jones, K. and Wilks, Y. (editors) (1983), *Automatic Natural Language Parsing*, Ellis Horwood, Chichester.
- Spärck Jones, K. and Willett, P. (editors) (1997), *Readings in Information Retrieval*, Morgan Kaufman, San Francisco.
- Tait, J.I. (editor) (2005), *Charting a New Course: Natural Language Processing and Information Retrieval. Essays in Honour of Karen Spärck Jones*, Springer, Dordrecht.
- Tuerk, A., Johnson, S.E., Jourlin, P., Spärck Jones, K. and Woodland, P.C. (2001) "The Cambridge University multimedia document retrieval demo system", *International Journal of Speech Technology*, Vol. 4, pp. 241-250.