Charting a New Course:
Natural Language Processing and Information Retrieval
Charting a New Course:
Natural Language Processing and Information Retrieval
Essays in Honour of Karen Spärck Jones

Edited by

John I. Tait
University of Sunderland,
School of Computing and Technology,
Sunderland, United Kingdom
Table of Contents

Preface, John I. Tait .................................................................vii

Karen Spärck Jones, Mark T. Maybury...........................................xi

A Retrospective View of Synonymy and Semantic Classification,  
Yorick A. Wilks and John I. Tait...................................................1

On the Early History of Evaluation in IR, Stephen Robertson...............13

The Emergence of Probabilistic Accounts of Information Retrieval  
C.J. van Rijsbergen ........................................................................23

Lovins Revisited, Martin F. Porter..................................................39

The History of IDF and its Influences on IR and Other Fields,  
Donna Harman................................................................................69

Beyond English Text: Multilingual and Multimedia Information Retrieval,  
Gareth J. F. Jones ............................................................................81

Karen Spärck Jones and Summarization, Mark T. Maybury...............99

Question Answering, Arthur W.S. Cater...........................................105

Noun Compounds Revisited, A. Copestake and E.J. Briscoe ...............129
Lexical Decomposition: For and Against, *Stephen G. Pulman* .........................155

The Importance of Focused Evaluations: a Case Study of TREC and DUC, *Donna Harman* ..................................................................................................................175

Mice from a Mountain: Reflections on Current Issues in Evaluation of Written Language Technology, *Robert Gaizauskas and Emma J. Barker* ........................................................................................................195

The Evaluation of Retrieval Effectiveness in Chemical Database Searching, *Peter Willett* ......................................................................................................................239

Unhappy Bedfellows: The Relationship of AI and IR, *Yorick A. Wilks* ..........255

Index ..................................................................................................................283
JOHN I. TAIT

PREFACE

I first met Karen Spärck Jones in 1977 in Chelmsford, Essex, England (of all places), at a seminar to which my undergraduate teachers Richard Bornat, Patrick J. Hayes and Bruce Anderson had invited me. It was a meeting which would change and enrich my life in many ways, and for which I will always be grateful to these men to whom I owe so much.

In April 1978 I began a Ph.D. under Karen’s supervision in Cambridge, cementing a relationship which continues to this day, and indeed is one of the reasons I ended up editing this volume.

I found Karen a highly stimulating supervisor and colleague, if not always easy to get along with. Indeed I believe some of our best work flowed precisely on those occasions when our discussions were, at the least, full and frank, disputatious and sometimes of extraordinary length. I can remember one occasion when Karen and I, and to an extent Bran Boguraev, spent over 6 hours arguing about an issue and in the end agreed to differ! From the areas about which we could agree there came one paper (Boguraev, Sparck Jones and Tait, 1982) from the areas about which we could not agree, Karen produced another paper (Sparck Jones, 1983), which I think stands as an excellent brief exposition on the subject (even if I still disagree with parts of its position). These two papers (I believe) were the first two computational linguistic papers from Cambridge on the subject of English compound nouns, starting a tradition of study which continues to this day, and indeed is reflected in the paper by Ann Copestake and Ted Briscoe in this volume.

Quite a number of years ago I decided that Karen’s contributions to the range of fields in which she has worked was such that a volume of this sort was appropriate, and that I should try to ensure it was produced. In 2001 I began sounding out various people to see whether there was sufficient support to make the production of a book viable. I was overwhelmed with the strength and range of support for the idea. Karen has always been a controversial and outspoken figure, but my concerns that that would over shadow other opinions and feeling about her were misplaced. Initially I had not intended or expected to be the editor, but Keith van Rijsbergen, in particular, took me to one side and made it clear he thought I should take on the challenge myself.
This initiated what has been something of an odyssey, over what turned out to be an extraordinarily difficult period in Karen’s life. First her own serious illness and then the untimely death of her husband Roger Needham cast a shade over the production of the book. Karen has bounced back from these difficulties in an exceptional and unique manner, further increasing my respect and admiration for her.

One of Karen’s great gifts is a profound intellectual rigour and ability to see when claims are not fully supported by experimental evidence or reasoned argument. She has made many major contributions over a range of fields which are represented in this book.

However it is sometimes not appreciated how well rounded a person Karen is. I remember her not only our academic work together, but also her interest in jewellery, in church architecture, and in sailing. The latter is what stimulated the title of the book, prompted by David Harper. Thank you, David for getting us onto a nautical course! The final form was a joint effort from Barbara Grosz, Keith van Rijsbergen and Yorick Wilks, so thanks are due to all of them, and indeed all those who contributed to the debate, as well.

It is perhaps worth noting the somewhat unusual approach we have adopted to Karen’s name. Wherever possible and sensible we have adopted the (probably more proper) form with a two-worded surname and an umlaut: Karen Spärck Jones. However, since it seems to continue to be the case that many automatic systems do not properly support umlauts, so in consultation with Karen we have generally used the form Karen Sparck Jones in references and the like in the belief that this will assist students and other scholars to find the referenced material.

The book is structured in five or six parts. Following Mark Maybury’s introduction and overview of Karen’s achievements, there are two chapters on early work (defined as prior to about 1975) from myself and Yorick Wilks, and Stephen Robertson. Next there are four chapters on Information Retrieval (IR) from Keith van Rijsbergen, Martin Porter, Donna Harman and Gareth Jones. This is followed by four chapters on Natural Language Processing (NLP) from Mark Maybury, Arthur Cater, Ann Copestake and Ted Briscoe, and Stephen Pulman. The third section covers one of the overarching themes of Karen’s career, evaluation, and contains two contributions (one from Donna Harman and one from Robert Gaizauskas and Emma Parker) which really draw together the themes of NLP and IR from the previous sections. Finally there are two chapters from Peter Willett and Yorick Wilks which relate Karen’s work to more recent broader developments, specifically chemoinformatics, and to Artificial Intelligence.

This brings me to one of my clearest conclusions from editing this book. Karen has produced a body of work which has never been so relevant as it is today. Despite the fact that some of it goes back over 40 years even the very early work can be studied with profit by those working on recently emerging topic, like the semantic web and bioinformatics. Indeed recently one of my own students who works on information discovery within the semantic web paradigm, after dismissing most of what has been done in IR since the 1970’s waved what turned out to be Karen’s thesis at me saying “but this is really worth reading”!
I would also like to thank Robbert van Berckelaer of Springer (formerly Kluwer) and Bruce Croft, the series editor, for their enthusiasm and unfailing support for the project during what has turned out to be a lengthy and tortuous journey.

Finally, I’d like to thank all those who have contributed to this volume in different ways, whether it be as authors, through reviewing or by giving encouragement and advice. At the risk of omitting someone important in addition to the authors I’d like to thank Ursula Martin, Queen Mary, University of London; Steve Pollitt, View-Based Systems Ltd; John Carroll, University of Sussex; Barbara Grosz, Harvard University; Wendy Hall, University of Southampton; Candy Sidner, Mitsubishi Electric Research Laboratories (MERL); Bonnie Webber, University of Edinburgh; Bran Boguraev, IBM TJ Watson Research Center. Hiyan Alshawi, Google; Bruce Croft, University Massachusetts Amherst; David Harper, Robert Gordon University Aberdeen.

John Tait, University of Sunderland.

REFERENCES


MARK T. MAYBURY

KAREN SPÄRCK JONES

Professor of Computers and Information (emeritus) and Computing

“Her energy and enthusiasm are truly legendary. Everyone who meets her is deeply impressed by the commitment and drive which she demonstrates in abundance. That this energy is backed by an exceptional intellect makes for a powerful combination.”

- Stephen Robertson, at the degree ceremony for Karen Spärck Jones’s honorary doctorate at City University

1. IN THE BEGINNING

From her early days, Karen exhibited signs of genuine intellectual curiosity. Her initial education was deep and broad - she read history and philosophy. As a teacher and researcher, Karen mastered knowledge of the past and created technology of the future. Karen’s extraordinary life as a computing pioneer is summarized in the timeline in Figure 1. Karen became involved with information retrieval (IR), in her own words, “for respectable intellectual reasons” but began serious IR work in the mid sixties through a “funding accident.” Working at the Cambridge Language Research Unit (CLRU) founded by the legendary Margaret Masterman was, according to Karen, “originally a lively discussion group interested in language and translation, subsequently funded to do research on automatic translation.”

From 1955 to 1958 the CLRU explored the value of Roget’s thesaurus within the context of machine translation and, subsequently, document retrieval. For translation, the headwords from Roget’s Thesaurus (325,000 words classified into 1073 headwords organized into 15 classes) were used for sense disambiguation during transfer so that, for example when the word “spend” occurred before “day” or “night” it was translated using its sense of “passing time” not dispersing money. For retrieval, Roget’s enabled word substitution so that, for example, a word like “eat” could be matched with “food”.

Karen’s doctoral research explored methods to exploit thesauri for language processing while fellow PhD student Roger Michael Needham was creating text-based methods for constructing retrieval thesauri (Masterman, Needham and Sparck Jones 1958). Roger Needham’s 1961 Cambridge PhD was on information retrieval, focusing in particular the application of digital computers to classification and grouping, using clump theory to classify archeological data (Needham and Sparck
Karen married Roger in 1958, they built a house, and in 1961 bought their first boat, later sailing round the east coast in an 1872-vintage Itchen Ferry Cutter.

2. DISSERTATION RESEARCH

Karen’s doctoral thesis (Sparck Jones 1964) was a cornerstone for research in the areas of synonymy and semantic classification. Karen proposed that a thesaurus could be constructed from sets of synonymous word senses derived by substitution in sentential contexts. Karen classified dictionary entries from the Oxford English Dictionary into semantic categories (headwords) found in Roget’s Thesaurus to create an ideal resource for language processing. This led later to a core idea that word classes could be derived by clustering based on lexical cooccurrence. Karen’s thesis was so novel and fundamental that twenty years later it was published as a book by Edinburgh University Press (Sparck Jones 1986). Yorick Wilks and John Tait present a “Retrospective of ‘Synonymy and Semantic Classification’” in Chapter One.

3. EARLY RESEARCH: A SCIENCE IS BORN

It is hard to imagine that when Karen came to the CRLU to do research on computing and language, there was not even one computer. Someone commented at the time that they were “Like children playing with invisible mice”. Yorick Wilks recalls “doing parsing with Hollerith card sorting machines”. Perhaps ironically CRLU staff did what has become, if now automated, a core element of modern language research methodology: corpus based processing. For example, at that time Mark Allford, who taught Russian and German, did some of the first corpus analysis in math and thermodynamics. He discovered that 1,000 terms would support reading 90% of text in mathematics whereas about 5,000 words were required for social sciences.

A rigorous scientist unsatisfied with only plausible arguments for ideas and seeking truth, using the Cranfield collection Karen began testing to see if classification would enhance recall in retrieval (i.e., return more relevant documents). Karen learned that collection frequency weighting of terms (aka inverse document frequency (IDF)) was cheap, useful, effective and applicable to many document collections. Donna Harman outlines in Chapter Five, the considerable influence of IDF on IR and NLP.

Because of the inability to predict the effectiveness of methods or explain performance, however, Karen also realized early on the many data variables and system parameters in indexing and retrieval systems demanded a finer descriptive and analytic framework. Early on detailed analytic experiments led her to suspect the argument that keyword clustering could enhance recall (Sparck Jones and Barber 1971), and she later discovered keyword clustering enhanced precision. Karen readily learned from others, exemplified by her adoption of Cornell’s approach to testing across collections. Seeking to address the challenges associated with
establishing collections to enable the scientific community to perform systematic comparative evaluations led Karen and colleagues to explore ideal test collections. When Keith van Rijsbergen returned from Australia to take up his Royal Society fellowship in Cambridge, Karen and Keith received a grant from the British Library to do a report on the ‘need and provision of an ideal test collection’. Keith gathered the data and worked out a preliminary design and authored a report on the need and provision of an ‘ideal’ test collection (Sparck Jones and van Rijsbergen 1975/6). This was followed by a report on a design study for the ‘ideal’ information retrieval test collection (Sparck Jones and Bates 1977). Finally, there was a report on the statistical bases of relevance assessment for the ‘ideal’ information retrieval testcollection (Gilbert and Sparck Jones 1979). The subsequent Text Retrieval and Evaluation Conference (TREC) was heavily influenced by the design for the ideal test collection. Donna Harman outlines Karen’s contributions to TREC in Chapter Eleven.

Karen Spärck Jones

Figure 1. Karen Spärck Jones Time Line

Karen recognized the key role of experimentation and evaluation (Sparck Jones 1986) in scientific progress for many tasks such as document classification,
retrieval, and translation. She embarked upon “a major series of comparative experiments” including larger scale tests both to understand indexing and retrieval but also to demonstrate generally applicable techniques. Karen reflected “I wanted to feel satisfied that the tests were valid, in being properly controlled and with performance properly measured. I believed that the standard of my own experiments, as well as those of others, needed to be raised, in particular in terms of collection size” (Sparck Jones 1988a; p. 4). Stephen Robertson elaborates on Karen’s early work on IR experiments and test collection design in Chapter Two.

Experimentation on collections led to “depression” at the lack of results from classification but “exhilaration” that collection term frequency weighting was useful and reliable. Rigorous and grueling experiments (many input, indexing, and output parameters, multiple test collections and request sets, and hundreds of runs on thousands of documents) ensured validity of results. One important discovery was the value of relevance weighting (Robertson and Sparck Jones 1976; Sparck Jones 1979a, Sparck Jones and Webster 1980) even with little relevance information (Sparck Jones 1979b). Karen credits influential examples set by Cyril Cleverdon, Mike Keen and Gerry Salton and notes collaborations with her colleagues (Keith van Rijsbergen, Stephen Robertson) and research assistants (Graham Bates and Chris Webster). Discouraged by the lack of “snap, crackle and pop” in IR research, she returned to the then more dynamic area of natural language processing (NLP).

4. INTELLIGENT KNOWLEDGE BASED SYSTEMS

Amidst her scientific contributions to IR and NLP, in 1984, Karen was instrumental in the establishment of the Intelligent Knowledge Based Systems (IKBS) research area, one of four key areas in the Alvey Program funded by the UK’s Science and Engineering Research Council (SERC). She proposed IKBS to the SERC which ultimately funded several hundred project teams drawn from industry, university, and polytechnics. An element of the program including creating instructional video lectures (the Alvey tapes) in areas such as logic programming, dealing with uncertainty, image understanding, machine learning, natural-language processing and of course expert systems, as well as starter kits with tools and instructional guides to help disseminate knowledge more broadly. The program also established a series of “community clubs” in a broad range of application areas including financial services (e.g., the Small Company Health Adviser), insurance, data processing (e.g., the Help Desk Adviser for data-processing installations), econometric modeling, real-time manufacturing planning, real-time quality control of processing plants, quantity surveying, transport-route planning, and water-industry construction planning. Khurshid Ahmad (University of Surrey) worked on an Alvey project at the time addressing knowledge based systems for water distribution network control and rehabilitation of sewer systems. Khurshid recounts:

Karen advanced the cause of information retrieval, especially knowledge-based IR, consistently, persistently, and laudably, throughout her career. Her views on the power and the limitations of intelligent systems are amongst most lucid ones: this was exemplified in a feature article on her and other leading intelligent systems
academics in the London-based Observer newspaper - a magazine article in which she was dressed as hiker or climber - during the Alvey Programme period.
Karen’s playful sense of humor was ever present.

Although an academic, Karen’s influence on the business community was also felt. Bob Moore of Microsoft Research recalls the mentorship Karen provided during 1985-1987 when he went to Cambridge to set up a research group for SRI International with an initial focus on natural-language processing. He notes:

At that time I was not many years out of grad school, and it was to be my first management position. I had only a vague understanding of the business side of contract research, and no knowledge at all about how things worked in the UK. Fortunately for me, Karen believed that expanding the base of NLP research in Cambridge by helping SRI was a Good Thing, so she virtually took me by the hand and guided me through the ins and outs of the local funding environment, including providing many introductions within the research departments of British companies, whose support we needed to provide matching funds to receive the Alvey grant.

5. NATURAL LANGUAGE FRONT ENDS TO DB

An early desire of artificial intelligence systems was to ease the human burden of access to not only unstructured sources like free text but also to structured sources like databases. Challenge language problems such as discourse (e.g., anaphora and ellipsis), sense ambiguity, indirect and partial language, and implicit statements, among others, vexed researchers. Karen and others recognized the importance of knowledge and inference in enhancing human natural language interaction with databases (Boguraev, Copestake and Sparck Jones 1986). This included the ability to detect and correct user presuppositions or misconceptions and to make inferences from their statements or the data itself (e.g., generalizations, deductions) to improve quality and naturalness.

6. NATURAL LANGUAGE ACCESS TO UTILITIES AND HETEROGENEOUS INFORMATION-INQUIRY SYSTEM

Karen and her collaborators investigated several rich forms of natural language access in the context of utility interfaces and inquiry systems. In the utility system case, in Menunet (Brooks and Sparck Jones 1985) users could access hierarchically organized menus by simply stating a particular action (e.g., “send” or “find”). The system would automatically construct a presentation of the various linguistic forms of accessible menus based on this input thus providing direct language access to a large number of possibly interrelated office automation functions.

In the second case of inquiry systems, language access was provided to a diverse set of information sources, from structured databases to unstructured text collections. Karen’s interest in this form of “question answering” included an early
recognition of the need for inference (Boguraev and Sparck Jones 1983, Sparck Jones and Tait 1984, Sparck Jones 1983, Boguraev et al. 1986). In order to bridge the gap between structured and unstructured sources, an unconventional approach was the use of shallow processing of knowledge. Unlike most conventional knowledge based systems, knowledge was viewed as representing relations among word senses rather than formally representing models of the world. Thus while inference was rather limited to such operations as establishing linguistic relations, substitutions, generalizations and refinements, it was also broader and potentially derivable automatically from on-line dictionaries or even unstructured document collections.

A natural extension of early work in natural language access to databases was the ability to question and get natural language answers and explanations from both structured and unstructured sources, such as the web. Arthur Cater addresses Question Answering in Chapter Eight.

7. TEXT SUMMARIZATION

The first collection of papers related to document summarization appeared in 1995 in a special issue on Text Summarization of the Journal of Information Processing and Management (Sparck Jones and Endres-Niggemeier 1995). This had roots in a 1993 Dagstuhl Seminar “Summarizing text for intelligent communication” in which Karen helped pull together an international group to focus on this important language application area.

In her characteristic style of crisply clearly characterizing the nature and scope of a problem, Karen articulated clearly the importance of distinguishing input factors (e.g., source form, subject type, unit), purpose factors (e.g., situation, audience, use), and output factors (e.g., material, format, style) with respect to summarization (Sparck Jones 1999). Early on, she forwarded important ideas such as the use of rhetorical structure and/or discourse purpose segmentation to enhance extraction of relevant units from source texts (Sparck Jones 1993).

8. MULTIMEDIA INFORMATION RETRIEVAL

Always pursuing novel directions, Karen in collaboration with others drew upon the foundations of IR and speech and language processing to move in the direction of retrieval of radio and television broadcasts. Her team was awarded a SIGIR best paper for their innovative research on the retrieval of spoken documents from multiple index sources (Jones, Foote, Sparck Jones and Young 1996) and the ACM Multimedia Conference best paper in the same year (Brown, Foote, Jones, Sparck Jones and Young 1996).

The Video Mail Retrieval Using Voice project developed robust unrestricted keyword spotting algorithms and adapted existing text-based information retrieval techniques to work effectively on voice and video data types. The Multimedia Document Retrieval (MDR) project (1997-2000) explored research into the audio
Indexing and retrieval of sources such as broadcast news (Johnson, Jourlin, Sparck Jones and Woodland 2001).

9. READINGS
Karen’s influence has not only been by her personal scientific contributions but her service to several research communities. Through her continuous, multi-decade service as a teacher, director and conductor of research, she influenced multiple generations of IR, speech and language scientists. She has not only enabled and shaped cadres of scientists and engineers, but she has also enhanced the quality and efficiency of the research community through the application of her significant knowledge and experience to the organization of the scientific literature. In addition to survey articles, she has collaborated to bring to life two instrumental readings: *Natural Language Processing* (Grosz, Sparck Jones and Webber 1986) and *Information Retrieval* (Sparck Jones and Willet 1997).

10. MPHIL IN COMPUTER SPEECH AND LANGUAGE PROCESSING
Not content with shaping the literature, Karen took an active and innovative role in graduate education in speech and language processing. Karen worked in the Computer Laboratory since 1968, but wherever she was she acted as a pioneer in research and an innovator in education. Collaborating with the late Prof. Frank Fallside from Engineering at Cambridge and Steven G. Pulman from the Computing Laboratory, Karen helped establish the first M. Phil. in Computer Speech and Language Processing at Cambridge in 1985 (renamed Computer Speech, Text, and Internet Technology in 2001). The interdisciplinary course was run jointly by the Engineering Department Speech Group and the Computer Laboratory Natural Language Processing Group with the assistance of the Department of Linguistics and the MRC Applied Psychology Unit. Consisting of two terms of lectures and practicals followed by a three month project and resulting thesis, the course introduced students to an interdisciplinary approach drawing students and lecturers from linguistics, psychology, computer science/artificial intelligence, engineering and mathematics. The program was distinguished by its balance of in-depth practical and theoretical grounding and strong links with industrial research laboratories.

Its course lecturers over the years (including Steve Young, Steve Pulman, Phillip Johnson-Laird, Ted Briscoe, Ann Copestake, Phil Woodland, Tony Robinson, Sarah Hawkins, Francis Nolan) and demonstrators were directly involved in leading edge research within the university and in collaboration with other European, US, and Japanese industrial and academic laboratories. Prof. Steve Young (University of Cambridge) notes that:
Over the time that Karen taught on the course there were approximately 320 graduates - many of whom are now well-known in either the speech or language field (or both) such as Michael Collins, Ted Gibson, Phil Woodland, Julian Odell, Philip Monaco, Tony Robinson, Simon King .... The course has grown in strength over the years with application rates now approaching 200 per year. To cater for the increased demand we recently increased the target number of places from 20 to 30.

In summary, Karen was instrumental in helping to establish an extraordinary educational team at Cambridge, now as a major international centre for education and research in speech and language processing.

11. ADVISOR

In addition to her active technical engagement with her research assistants in the multiple areas of endeavor noted above, Karen’s mark is seen in the lives of dozens of researchers who have gone on to influence numerous scientific communities, from information retrieval, to intelligent agents, to automated summarization, to language generation. Table 1 shows Karen’s Ph.D. students, whose topics span the broad areas of information retrieval, machine translation, expert systems, user modeling, and natural language processing, reflecting Karen’s wide ranging interests and influence.

12. PROFESSIONAL SERVICE AND HONORS

Throughout the years Karen has served her professional community in multiple capacities including as a Member of the Foresight Programme Panel on Information Technology, the Committee on Linguistics in Documentation, the Executive Committee of the Museums Documentation Association, the Department of Philosophy Advisory Board at Carnegie Mellon University, and the Advisory Committee for the Research and Innovation Centre of the British Library.
Table 1. Karen Spärck Jones Ph D students (partial)

<table>
<thead>
<tr>
<th>Doctoral Candidate</th>
<th>Subject</th>
<th>Start</th>
<th>Finish</th>
<th>UCAM TR #</th>
</tr>
</thead>
<tbody>
<tr>
<td>Martin Porter</td>
<td></td>
<td>1967</td>
<td>1969</td>
<td></td>
</tr>
<tr>
<td>Branimir Konstatinov Boguraev</td>
<td>Automatic resolution of linguistic ambiguities</td>
<td>1979</td>
<td></td>
<td>TR-11</td>
</tr>
<tr>
<td>Hiyan Alshawi</td>
<td>Memory and context mechanisms for automatic text processing</td>
<td>1984</td>
<td></td>
<td>TR-60</td>
</tr>
<tr>
<td>David Carter</td>
<td>Interpreting Anaphors in Natural Language Text</td>
<td>1986</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Victor Poznanski</td>
<td>A relevance-based utterance processing system</td>
<td>1985</td>
<td>1990</td>
<td>TR-246</td>
</tr>
<tr>
<td>Derek G. Bridge</td>
<td>Computing presuppositions in an incremental language processing system</td>
<td>1986</td>
<td>1991</td>
<td>TR-237</td>
</tr>
<tr>
<td>Richard C. Hutchings</td>
<td>Natural language processing</td>
<td>1986</td>
<td>1991</td>
<td></td>
</tr>
<tr>
<td>Richard I. Tucker</td>
<td>Automatic summarising and the CLASP system</td>
<td>1993</td>
<td>1999</td>
<td>TR-484</td>
</tr>
<tr>
<td>Oi Yee (Olivia) Kwong</td>
<td>Word sense selection in texts: an integrated model</td>
<td>1996</td>
<td>2000</td>
<td>TR-504</td>
</tr>
<tr>
<td>Martin Choquette</td>
<td>Automatic information retrieval</td>
<td>1996</td>
<td>2001</td>
<td></td>
</tr>
<tr>
<td>Donbla B. Nic Gearailt</td>
<td>Natural language processing</td>
<td>1997</td>
<td>2002</td>
<td></td>
</tr>
</tbody>
</table>

In 1988, Karen was awarded the Gerard Salton Award from the ACM Special Interest Group in Information Retrieval (SIGIR) for her research achievement. Karen served as Vice President of the Association for Computational Linguistics in 1993 and then became President in 1994. She has been a member of the DARPA/NIST Text Retrieval Conferences Program Committee since 1994, and of the DARPA Translingual Information Extraction Detection and Summarization (TIDES) Program Advisory Committee since 1999. In 1993 Karen became a Fellow of the American Association of Artificial Intelligence (AAAI) “for contributions to
applied natural language research; leadership of an internationally renowned natural language processing research group; and work as an educator.” In 1999 she became a Fellow of European Coordinating Committee for Artificial Intelligence (ECCAI) for making significant and sustained contributions to artificial intelligence. In 1995 she was elected a Fellow of the British Academy. And in 1997 Karen Ida Boal Spärck Jones gained admission to the degree of Doctor of Science, Honoris Causa, from City University. Then, in 2002, Karen gave the prestigious Grace Hopper Lecture which serves “the dual purpose of recognizing successful women in engineering and of inspiring students to achieve at the highest level”. Characteristic of her enthusiasm and drive, the final words on the slides to her talk “Language and Information: Old Ideas, New Achievements” (Sparck Jones 2002a) regarding statistical language and information processing read “GO FOR IT!” Topping off this impressive list of accolades, on July 23rd, 2004 in Barcelona, Spain Karen Spärck Jones received the ACL’s Lifetime Achievement Award.

13. KAREN THE PERSON

Those who had the honor of learning from and with Karen immediately recognize her challenging but generous approach. She demanded clarity in communication, coherence and cohesion in writing, and evidence and persuasiveness of argument. Rob Gaizauskas (University of Sheffield) recalls being terrified by his first meeting with Karen but soon understood her “intensity and genuine concern for others doing a good job”, benefiting from her academic guidance in the now widespread GATE. Ralph Wesichedel (BBN) has always been impressed with Karen’s “passion for excellence, desire to bring out the best in others, and her commitment to evaluation progress”.

Karen possesses an uncanny capacity to use all of her senses and apply her active scientific mind to exploit the environment around her to discovery for new knowledge. For example, Jeremy Bennett (University of Bath, England) reported how Karen, a contributor to the analysis of compound nominals, discovered a school signpost visible when driving from Cambridge to Newmarket, UK containing eight nouns in a row (a record?): “The Horse Race Betting Levy Board Apprentice Training School”. Ann Copestake and Ted Briscoe describe noun compounds in Chapter Nine.

Another notable characteristic of Karen is her rich use of metaphor in her principally scientific writing. For example, in her writings she warned of techniques that are simply “bombinating in the void” or of the “rush to climb on the information bandwagon.” She argued for solid scientific progress as opposed to “just inventing copy for the salesman”. She described how the “the library schools [will] train the professionals to guard the sacred flame of bibliographic control”. And she wrote how we need to “drive useful roads through the enormous jungle we are currently just viewing on day trips from the outside”.

Her well-developed sense of humor is evident throughout her writings. In the introduction to the Readings in NLP, for example, she notes “There is a widespread belief that AI-oriented NLP research began in the late sixties, and an equally
KAREN SPÄRCK JONES

widespread myth that the early workers in machine translation were all crooks or bozos” (p. xiii)

Throughout the years, Karen was an inspiration to her colleagues. For example, Professor Yorick Wilks (University of Sheffield), states that:

the most remarkable thing about Karen is that, as is now clear many years later, she pioneered a whole field of statistical, data-driven semantics, but, because of the vicissitudes of the times and the [CLRU] laboratory where she worked, she could not publish or continue to work in it, so she went off and founded a very distinguished career in something completely different, only returning to her early interests many years later. I think this is unique in the history of our field.

A comment echoed both others, Candy Sidner notes her admiration for Karen throughout the years. Ed Hovy (USC/ISI) calls Karen the “éminence grise” describing her as “the voice or reason and historian of the field.”

14. THE FUTURE

We are all blessed that Karen remains an active and influential language scientist and engineer. This is more poignant today with the recently published Computer Systems: Theory, Technology and Applications: A Tribute to Roger Needham (Herbert and Sparck Jones 2004), a collection tribute to the tragic passing of Karen’s husband. If the past is any indicator, Karen will likely surprise us again. A laypersons guide to security (Sparck Jones 2002b) and journal article on privacy (Sparck Jones,2003b) are just the latest installments. Past performance suggests she will apply her formidable talents to a practical application area of language process that has some core scientific challenge. We look forward to Karen continuing to spark action in herself and others!

Mark Maybury, MITRE Corporation

ACKNOWLEDGEMENTS

I thank Keith van Rijsbergen (University of Glasgow) and Yorick Wilks (University of Sheffield) for their corrections on and contributions to earlier drafts.

REFERENCES

A RETROSPECTIVE VIEW OF SYNONYMY AND SEMANTIC CLASSIFICATION

1. INTRODUCTION
Karen Spärck Jones' Cambridge PhD thesis of 1964 has had an interesting and unusual history. Entitled Synonymy and Semantic Classification (henceforth SSC) it was reproduced only in the simple mimeo book form then used by the Cambridge Language Research Unit where she worked. It was finally published in 1986, in an Edinburgh University Press series. Even that late publication managed to be ahead of a great deal of later work that recapitulates aspects of it, usually from ignorance of its existence. There is no doubt that SSC was developing statistical and symbolic techniques for the use of what we now call language resources so far ahead of other work that it was almost impossible for contemporary researchers to understand the book or to relate it to their own activity. At the time SSC was being written, Olney and Revard (1968) were exploring the content of Webster's Third Dictionary quantitatively on punched cards at Systems Development Corporation (where Sparck Jones also was by chance, joining in their work during 1967) and their work met a similar lack of reception, it, too, being twenty to thirty years ahead of its time.

2. A BRIEF OVERVIEW OF SYNONYMY AND SEMANTIC CLASSIFICATION
SSC begins with a review of the implications of the use of the computer as a tool to study natural language text. It discusses the need for precision of representation (in dictionaries, grammars and thesauri) for automatic processing, but rapidly moves to a deeper discussion of meaning, focussing in particular on the claim that in the context of a coherent text (fragment) different words will be used in senses with related meanings. There is a developed example in which canal and road are cited as means of communication. Roget's thesaurus is then put forward as a means of operationalising this intuition. The approach adopted finds strong echoes in much later work by Morris and Hirst (1991) and Ellman and Tait (2000), which do not really share the same intellectual heritage.

Sparck Jones then moves on to consider the notion of semantic relations between words: her focus is on synonymy and on Lyons (1961), although a range of other

---

1 SSC page references refer to the Edinburgh edition.
relations (including antonymy, hyponymy, logical implication, Lyon’s incompatibility and so on) and a range of other authors are mentioned. Chapter Two ends with a proposal to test a notion of synonymy based on substitution, reducing a subtle and complex notion to an empirically testable notion, without losing sight of the limitations of the test.

Chapter Three and Four develop this notion of synonymy based on substitution by using the notion of a row or set of close synonyms (cf Wordnet synsets). The discussion is sophisticated in many ways, but suffers from the use of an obscure notion of a play (a kind of semantic interpretation), from considering the context of the use of a word only in terms of the sentence (and not more broadly), and from the attempt to move between a specific word-use and a word-sign (string of characters) without any intermediate notion of morphology (strictly graphology) or intermediate word senses. This is not to say that taking on these notions would necessarily simplify the discussion: but they sometimes make the discussion hard to follow to at least one of the current authors’ eyes. The chapter moves on to a fascinating discussion of a notion of semantic distance (likeness) between words (and then between phrases) based on similarity of their occurrence patterns in thesaurus rows.

Chapter Six describes a series of Practical Experiments, using an analysis of Richards’ book “English through Pictures” which reports some success in building a simple prototype system of the kind described in the previous chapters.

The thesis concludes with some manual experiments concerning the feasibility of discovering the semantic relationships between words in coherent text and then argues that, taken together, these experiments support the notion that there is conceptual repetition in discourse (p 200).

A brief summary like this is inevitably unfair to the original. Some passages, even now, reveal a deep understanding of aspects of language which we have yet to fully face up to in Computational Linguistics (CL). For example, some of the discussion of metaphor in Chapter Two and Three, and the kind of conflict between specific use and overtones of a word derived from its whole range of uses (SSC, p86) shows great sophistication.

We now return to the strengths and weaknesses of SSC which we outlined at the beginning of this section. We will then pass on to highlight some aspects of this work which resonate with more recent developments in Computation Linguistics and Information Retrieval, despite the fact that it is now over forty years old.

3. STRENGTHS AND WEAKNESSES OF SSC

SSC has three great strengths. First, SSC brought together Information Retrieval (IR) methods with linguistic semantics and CL for the first time, a link that is now accepted and productive (as well as the subject of her 1999 AIJ article (Sparck Jones 1999), and thus an interest spanning her career). In saying that, we do not imply SSC is about IR, but that the underlying clustering algorithm she applied in it to thesaurus rows was the so called Theory of Clumps (1962) of Needham and Parker-Rhodes, a development in automated classification of Tanimoto’s (1958) original idea for derived clusters as a basis of IR. The principal originality of SSC was to take an IR
SYNONYM AND SEMANTIC CLASSIFICATION

clustering algorithm and apply it to features that described not physical objects or documents but other words or features at the same level as the classifiers themselves, and to which they were bound by a defined relationship of semi-synonymy. The kinds of associative nets/clumps she derived have been rediscovered many times since by others, probably in part because her thesis was not published e.g. Schvaneveldt's Pathfinder networks (1990) which were patented for IR.

Secondly, SSC's use of Roget's Thesaurus is possibly the first use of an established machine-readable linguistic/lexical resource in CL, apart perhaps from the roughly contemporary quantitative computations with Websters' Third Dictionary by Olney and Revard at SDC mentioned above. The widespread use of linguistic resources, such as machine-readable dictionaries, as a basis for NLP did not become commonplace until the late Eighties, when among the earliest contributors were members of her own laboratory, such as Bran Boguraev (Boguraev and Briscoe, 1989).

Thirdly, SSC shows an appreciation of the need to evaluate ideas about language processing by experiments on realistic samples of language using well-defined tasks, a matter we now take for granted but, when SSC was written, Artificial Intelligence (AI) was still in the heyday of its toy systems with tiny sets of examples.

However, the principles underlying SSC as well as its implementation and evaluation, unfair as it perhaps is to raise these modern notions for work done 40 years ago, still do give rise to real problems and we set out some problems with SSC that were always evident and have not changed with time (as its virtues have in the list above).

There are serious shortcomings in the discussion of the experiments which are very hard to interpret: there is a lack of detail (for example algorithmic descriptions) preventing judgements being made about the scalability of the algorithms. There is also a lack of clarity about the experimental set ups: some are clearly manual, some apparently automatic, one probably semi-automatic. The general notion underlying the experiments is very clear: namely, applying the theory of clumps to features, where in SSC the features were words whose features were being-corow-members, which should have resulted in clumps of words associated by the clump algorithm. But the matrix inversions required for that computation were very large and almost certainly not tractable over a database the size of Roget. The whole of Roget's thesaurus was put onto punch cards by Betty May, but only a sample can have been used in the experiments described in SSC. Chapter 6 note 14 clearly implies the adaption of the ideas to the practicalities of computing with then available machines.

One of the problems in interpreting SSC today is confusion between what was achieved with the then available computing engines and knowledge of software engineering, and what could have been achieved if Sparck Jones had had today's computers and software engineering and further was doing this work with the benefit of the insights on language we have gained over the intervening 40+ years.

There is also a failure to grasp the problems posed by basing the synonymy analysis on the use of words in context, presumably the meaning of this word in this sentence (for simplicity let us confine ourselves to writing) in this text at this time to this reader, as opposed to a word sense in a dictionary. Indeed, there is a slide between the two, with the definition of word-use on page 79 being concerned with
“ployed” sentences versus page 122 where word-use is defined by existence in a row. The discussion in Chapter 4 seems to show Sparck Jones is aware of the problem, but she shies away from the introduction of an intermediate layer between word-use and word-sign, one that corresponds to what we all now refer to, without much sign of scepticism, as a word-sense. One might (then or even now) put forward the objection that this is introducing an artificial abstract notion into a system which is otherwise entirely dependent on directly observable phenomena in language. But might not the avoidance of this (conventional) abstraction be the reason the system has the problems it undoubtedly has, dealing with the more complex relations like hyponymy or antonymy?

These last arise from a definition of likeness (SSC p102) which seems to gloss over the previous distinction between word-use and signs. Sparck Jones is clearly aware that there are complex relations between notions of substitution, hyponymy, homonymy, and synonymy, but only the first is given an operational definition with any plausibility.

In modern terms, there may be a parameter of the machine learning algorithm in which every occurrence of every sign (to use SSC's terminology) has its own row (cf. SSC p90) and a much smaller collection of rows emerge, but there is a danger of hyponyms and synonyms occurring in the same row. Antonymy, too, must be part of word meaning but the structure seems unable to take account of this.

The output from unsupervised methods is notoriously hard to interpret: given clumps of row-associated words from the program, why would they be better clumps than those provided by the Thesaurus heads themselves? No answer to this could have been expected at the time, and is barely available now: there is an awareness in Chapter 7 of the need to measure this output against some operational task, such as machine translation, though that was of course beyond the scope of SSC. The basis of the property of co-row-ness (for words) is that of substitution-preserving-some-property: Sparck Jones discusses this notion and its evident circularity yet she goes on to adopt it and then identify that with Thesaurus rows. She refers to, and is clearly aware of, Quine's critique of any such notion as circular (1953). There is a double sleight of hand here: even if substitution does provide a test of rowness, why should we accept Roget's rows as passing it, as she clearly does in order to get a data set? One could say that SSC's rows are ambiguous between an emergent property of language use (corresponding to unsupervised learning in more modern usage, and Parker-Rhodes & Needham's clumping) and artefacts extracted from a human constructed resource: such as Roget's rows, and, later, Miller's (2000) synsets (corresponding to supervised learning perhaps). This is perhaps best illustrated by considering the question: if the practical experiments of Chapter Six had produced row systems quite unlike Roget, what would this have meant for the (implicit) hypothesis of the whole thesis?

SSC is presented explicitly as a search for emergent semantic primitives; but how do (or could) these emerge from these computations? Yet, by using Roget she already assumes such a set (the 1000 heads of Roget): so why is that set worse (or better) than any she derives, or might derive with further computation? Perhaps what is lacking, in modern terms, is an understanding of the need for an objective
function, allowing us to distinguish more and less optimal solutions the need for which is now so well understood in unsupervised machine learning.

This might seem a long list of problems. However in view of the groundbreaking nature of the work, the intellectual tradition from which it sprang, the extraordinarily limited computational environment in which it was undertaken, they are comparatively minor, and in no way detract from the major strengths.

4. A FURTHER TWENTY YEARS LATER: SPARCK JONES'S VIEW OF SSC AFTER 20 YEARS.

Sparck Jones wrote a new introduction to SSC when it was finally published, 20 years late, in the Edinburgh University IT series, run by one of the present authors. Perhaps the most striking feature of her retrospective, as compared to the original SSC, is the emphasis on semantic primitives and the explicit opening claim that “The thesis proposes a characterisation of, and a basis for deriving, semantic primitives, i.e. the general concepts under which natural language words and messages are categorized (p.1)”.

This view of SSC is not one that a reader of the original thesis would necessarily come to from its text, although it makes perfect sense if we take semantic primitives to mean the topic markers that are the 1000 or so Roget heads, such as 324 SOFTNESS. However, and as noted in the previous section there are some problems with reconciling this notion of predefined primitives and truly emergent ones. In her retrospective discussion Sparck Jones widens comparisons at this point, describing such primitives as domain dependent (e.g. SHIP-as-a-type) by contrast with more general notions of semantic primitive in the work of Katz (1972), Wilks (1975) and Schank (1975), and which was criticised by Lewis (1970), Pulman (1983) and others.

These latter primitives (usually equivalent to notions such as human, physical object, movement etc.) she takes as being general rather than domain dependent, which suggests the two types could all be fitted together in some very semantic hierarchy with physical object near the very top and types of ship at the bottom; and this is something like what one gets in Wordnet and indeed in the hierarchy Roget himself offers at the start of his Thesaurus. That Sparck Jones sees these two types of primitive as closely related, as is shown by the original appendix to SSC on Thesauri and Synonym dictionaries , a historical excursus that covers both types of primitive and remains for some the best thing in the book.

In the structures associated with the LDOCE dictionary (Procter, 1978) and both types are given as quite separate hierarchies (of semantic and domain terms) and dictionary entries are decorated with both as features independently. Again, in much recent work on word-sense disambiguation (e.g. Yarowsky (2000), Wilks and Stevenson (1997)) both types of hierarchy have been used as separate information sources, combined ultimately by the algorithm, but where it can be seen that one type tends to disambiguate verbs and the other nouns. None of these considerations are definitive as to whether there are two levels or types of primitives or not, or whether the difference is merely one of degree and domain. Sparck Jones certainly
distinguishes two roles for primitives, as do many authors, namely being definitional of sense (as in a dictionary) and being selective for particular senses (as in a disambiguation program) but that distinction has no implications for the one- or two levels of primitive issue.

5. THE VIEW OF SEMANTICS EMBODIED IN SSC

As noted above, SSC was perhaps the first attempt to capture computationally the elusive notion of linguistic relations or fields, one well established in the descriptive literature (Lyons, 1961) but with no formal or computational basis up to that point. It is notion close to some Continental notions of text structure and meaning, ones that have received wide popular discussion, and in which the meaning of any symbol depends, by a relation of contrast, on its relation to other symbols, rather than to objects in the world, as in the basic, rather simple minded, version of Anglo-Saxon philosophy.

Of course linguistic or semantic fields are a subtle and complex subject. A later review (Lyons, 1977, Chapter 8) points out some commonalities, but also contradictions and contrasts between different field theories. In the most accessible form of the theory, there is postulated some sort of meaning surface lying between the lexemes of a language and the world of language use. Particular lexemes are then related to areas of this meaning surface. Most field theorists are concerned with changes in the meaning of language over time, and this creates an odd contrast with SSC, which, like almost all computational work which followed it, takes a rather static, or at least snap shot, view of language. As we have noted, all field theorists share a focus on lexical semantics, in terms of the relations between words and other words or the whole vocabulary, which is presumably what made the approach attractive to Sparck Jones, but they also share a difficulty in formalising the notion of field in a consistent and useful way.

Much of the discussion of SSC shies away from putting forward anything which cannot be directly observed in text. In the end SSC resorts to concepts as additional, artificial, constructs lying outside observable language. One might say the work is caught between Skinner and Saussure, having on the one side the poverty of sticking to the merely observable and on the other problem of subjecting the abstract to empirical verification. In a later overview of her work in IR (Sparck Jones, 2003) she refers to a simple principle underlying everything she does as “taking words as they stand” a position already present in SSC, before Sparck Jones began her distinguished career in IR, namely a reluctance to decorate words with logical, primitive semantic, or other linguistic codings (as opposed to relations). This was something shared, in an interdisciplinary way, with linguistic field theorists and their Continental counterparts. Against this, it could be argued that, by accepting, as she did, the overarching a priori architecture of Roget, all derived from a single mind by intuition, Sparck Jones was accepting a great deal of decoration beyond the words themselves. Conversely, it can be argued, changing sides as it were, that nothing violates that principle in using a thesaurus or a dictionary because the decorations are only more words, as are the Thesaurus heads of Roget themselves.
SYNONYM AND SEMANTIC CLASSIFICATION

6. OTHER RESONANCES BETWEEN SSC AND MORE MODERN WORK IN MT, NLP AND IR.

The discussion of the likeness between words and phrases in Chapter Three of SSC, referred to in section 2 above as a form of semantic distance, finds many echoes in later query expansion techniques, like pseudo-relevance feedback or local feedback (Xu and Croft, 1996). These techniques presume that terms which co-occur in documents with query terms are semantically related to query term uses. They rely on the implicit existence of an empirically derived thesaurus, or clump dictionary, on which similarity calculations of the sort described in SSC can be computed.

The introductory material to Chapter Five contains a couple of oddities which hide really quite deep insights.

First, there is discussion of the very large number of rows in which a word might be placed reflecting the very fined grained distinctions of sense which might be required for high-quality machine translation. However, oddly, there is no discussion of how one might link these to another natural language. Was Sparck Jones perhaps thinking that some form of parallel corpora would solve this problem given the automatic procedure? Or was the problems posed by the need to link the source and target languages simply missed?

Secondly, initially she proposes to distinguish every sentence position of every use of a word, but this abandoned on grounds of efficiency. However retaining this position would imply the learning not only of a synonym dictionary but also of a corresponding grammar in some sense. Further it might imply a finite model of language (in the absence of a generative component). It is hard to believe these restrictions were an oversight in view of the sophistication of the discussion elsewhere. Sparck Jones and her collaborators clearly understood such a process might imply learning or deriving a grammar stored in the thesaurus (Masterman, Needham and Sparck Jones 1958), but perhaps not its implications for the underlying model of language.

7. WHAT WAS THE SSC COMPUTATION/ALGORITHM?

It is clear that Sparck Jones in SSC made use of the Theory of Clumps, an unsupervised classification algorithm, deriving ultimately from Tanimoto and refined by Roger Needham (her husband) and Frederick Parker-Rhodes at CLR.U. The Theory of Clumps (from now on TC) which she found ultimately unsatisfactory for her purposes (see the quotation above) was an algorithm that took a set of objects $x$ classified by a set of features $y$ and produced clumps or sets drawn from $x$ which expressed natural subsets of $x$ in terms of the assigned features. An aspect of TC which Sparck Jones liked and drew attention to (as did Roger Needham) was that it had a feature close of Wittgenstein’s notion of family resemblances namely that subsets, so found, did not need to share any common feature at all and hence this notion was not at all part of the old Necessary and Sufficient Conditions tradition for being a thing of a certain sort.

Roger Needham’s thesis was classic application of TC, outside IR that is, and he took a set of Greek pots classified by a range of features (colour, handles, decorative
figures etc.) and produced plausible sets of pots based on the core notion of TC that things should be seen as alike if they tended to have the same features, or had separately the features that other things had as common features etc. It was thus an associative rather than definitional model of similarity and would have fallen under Firth’s phrase about words and “knowing them by the company they keep”. Things in the same clump would tend to keep the same company in terms of features.

Sparck Jones’s application of TC was thus more original than taking objects and features as quite different sorts of thing: she realised that both could be words and that words as features could be used to classify words as objects. Thus her classification relationship was that of appearing in the same row in Roget’s thesaurus. Elsewhere in this paper we discuss the implications of that assumption of classification as a form of synonymy but here we simply note that, in TC terms, co-row words were features of any given member word, where the co-row members were derived from the OED by seeking in entries for semi-synonyms and testing their substitutability (intuitively) within the example sentences given in the dictionary.

Given this assumption, TC could proceed, which meant first a matrix of features against objects was constructed, notionally at least, and here, since the matrix is symmetrical (both sides being in principle the whole vocabulary of Roget words) we can imagine a matrix with something of the order of 50K rows and columns. At this point forms of the TC algorithms come into play, of which the most basic is a measure of how close any two rows (derived as above) are. Sparck Jones adopts a rough and ready measure of the number of common words in two rows divided by the total number of distinct words in both i.e. their intersection divided by their union.

The main TC algorithm then runs and produces tentative clumps of objects based on the object-feature associations established in this way. These clumps should, being empirically based on associations in a corpus (the OED) yield better groups than Roget heads (considered as groupings of semi-synonyms). On p.183 she writes of assuming that we now have a better thesaurus than Roget’s, but one of the same kind, and one that might be tested against Roget in simple Machine Translation (MT) experiments.

This shows clearly that the clump output from SSC was of the same type as Roget heads themselves and, at one point, she discusses a possible recursive procedure for organizing the clumps produced by the program into a flat hierarchy more like Roget itself.

The account above must be treated with caution because of the different way experiments were handled and described then and now, and in large measure because, as we noted earlier, the 50K square matrix could not be constructed with the computers then available, nor were there as many techniques then for representing large very sparse matrices in alternative, more compressed, forms. Hence whatever experiments she did were necessarily on very small samples and she shows a sample of 500 rows and describes an experiment based on 180 rows (p.170), the maximum number her program could handle. An output clump is given as a set of rows deemed sufficiently close (p.172) and there is an extensive discussion (pp.176-181) of how this should be evaluated (by comparing it with