ACL Lifetime Achievement Award

Some Points in a Time[1]

Karen Spärck Jones*
University of Cambridge

This article offers a personal perspective on the development of language and information processing over the last half century, focusing on the use of statistical methods. Introduced, with computers, in the 1950s, these have not always been highly regarded, but were revived in the 1990s. They have proved effective in more ways than might have been expected, and encourage new thinking about what language and information processing involve.

First, to say how much I appreciate the completely unexpected honour of this award, and to thank the ACL for it.

I want to look at one line, or thread, in natural language processing (NLP) research: how it began, what happened to it, what it suggests we need to investigate now. I shall take some papers of my own as pegs to hang the story on, but without making any claims for particular merit in these papers.

My first paper, "The analogy between MT and IR," with Margaret Masterman and Roger Needham, was for a conference in 1958. The analogy it proclaimed was in the need for a thesaurus, that is, a semantic classification. Machine translation (MT) and information retrieval (IR) are different tasks in their granularity and in the role of syntax. But we argued that both need a means of finding common concepts behind surface words, so we at once identify text content and word senses. We took Roget's *Thesaurus* as such a tool in our MT experiments (with punched cards), where we exploited the redundancy that text always has to select class labels, that indicate senses, for words.

The essential idea is illustrated, in extremely simple form (as with all my examples) in Figure 1. If we have to translate *The farmer cultivates the field,* where *field* has a range of senses including LAND and SUBJECT that may well have different target language equivalents, the fact that the general concept AGRICULTURE underlies each of *farmer*, *cultivates,* and *field* selects the sense LAND for *field*.

But we found in our research that existing thesauruses, like Roget's, were not wholly satisfactory, for example through missing senses; and we wanted to build a better one, ideally automatically. The natural way to do this, the obverse of the way the thesaurus would be applied once it was constructed, was by using text distribution data for words and applying statistical classification methods to these data.

There were of course no corpora available then, so in my thesis work I finessed getting the input data for classification by taking dictionary definitions, which often consist of close synonym sets, as showing the most primitive and minimal shared-

* Computer Laboratory, William Gates Building, JJ Thomson Avenue, Cambridge CB3 0FD, UK. E-mail: ksj@cl.cam.ac.uk.

1 This article is the text of the talk given on receipt of the ACL's Lifetime Achievement Award in 2004. The figures provided in the article reproduce the slides used for the talk.

M. Masterman, R.M. Needham and K. Sparck Jones
The analogy between mechanical translation and library retrieval, 1959

thesaurus as semantic classification
class labels as interlingua for
index and match in retrieval (Luhn)
select and replace in translation

The farmer cultivates the field
field = LAND/REGION/SPHERE/SUBJECT
AGRICULTURE repeats select *field* = LAND

Figure 1

context behavior for the words concerned. The words in such a set could be taken as mutually substitutible in some context, which could be checked by reference to the text examples given in the dictionary. This mimicked what could in principle be delivered by a corpus which showed one word or another in the set occurring in the same text context. I then showed that one could apply general classification methods, notably Roger Needham's theory of clumps, to find larger classes of words with similar behavior.

For example, suppose we have synonym sets, or "rows," as shown in Figure 2, that is, *activity animation movement, activity movement business, briskness business liveliness*, etc., each representing substitutive, i.e., synonymous, *uses* of the words concerned. We can then derive classes of words with similar uses by exploiting the recurrence of the same word symbols across different rows: The senses of a word in different rows are distinct senses, but they can legitimately be assumed to be semantically related simply through the fact that they are uses of the same word sign.

In subsequent work, Kenneth Harper and I showed that it was possible to derive classes from actual word occurrences and cooccurences in text, as illustrated in Figure 3. Given a (hand-) parsed Russian scientific text corpus, we investigated the collocation behavior of 40 Russian nouns (here represented by English equivalents) in governor/dependent relations, and grouped them by their shared collocates. Somewhat suprisingly, for such a small corpus, we were able to extract semantically

Synonymy and Semantic Classification, 1964/1986

dictionary definition equivalents taken as
sharing text contexts
group by similar contextual behaviour

500 equivalence sets, clumping:

activity animation movement
activity movement business
business briskness liveliness

$\Longrightarrow$ *activity animation movement business*
briskness liveliness

Figure 2

A small semantic classification experiment using cooccurrence data, 1967

nouns with governor/dependent collocates
group by shared collocates

40 Russian nouns, 120 K words parsed text:

- *height depth width*
- *question problem2*
- *atom gas ion copper metal proton silver uranium*
- *calculation1 measurement study investigation determination ratio consideration calculation2 comparison*

Figure 3

plausible classes, like *height depth width* or *calculation measurement study investigation* etc. as shown in Figure 3.

These tests were very small in absolute terms (though less so relative to quite recent research than might be expected). But the much more serious problem, for evaluation, was that there was no MT system you could plug a classification into to test it. Different general classification methods can give different, but equally plausible, classifications, so you need an application context to choose among them, as well as to check that the generic idea of such a sense selection apparatus is sound.

We also needed a theory of discourse structure, above the sentence, to support and constrain the use of a classification. MT then was sentence-limited, but in general you cannot rely on individual sentences for all the information you need for sense resolution. Figure 4 shows a very rudimentary attempt to model discourse structure using topic/comment relations across sentences, shown using T or C with numbers for concepts. Topic/content linkage patterns could show where to look for information from other words to help select the sense of a word in a particular location. For instance, in selecting a sense for a given topic word, concepts that were shared with previous topic words, or recent comment words, would be preferred. In the illustration the same word is repeated, but it is easy to imagine, for example, that *ocean* could occur at one point and *sea* at another, so selection would involve semantic classes.

Notes on semantic discourse structure, 1967

semantic classes and discourse patterns
discovering message structure
constraining sense selection

The Pacific is by far the world's largest ocean.	1T 2C
Scattered over it are thousands of islands.	1T 3C
These islands make stepping stones across the ocean.	3T 4C 2C

Figure 4

But work like this was only the sketchiest beginning as far as MT was concerned. Retrieval was a much better initial test field for semantic classification, because you can get index keys for documents without interference from the need for sentence parses. At the same time, you have to consider the document as a whole. Retrieval, for the growing technical literature, was also a very pressing practical task, so people were developing evaluation methods for different indexing and searching strategies. Researchers wanted to show that a statistical classification, based on term distributions across documents, was effective as a lexical substitution device, helping recall (an idea first published by Luhn [1957]).

The task function for classification in IR was important because it forced thinking about the choice of classification model, for example should classes be overlapping or exclusive? Deriving classes is a major step beyond computing word pair associations. It needs sound and appropriate, as well as computable, approaches, desiderata that are not always recognized even now.

We began to test automatic classification for retrieval with a small collection built by Cyril Cleverdon and also used by Gerard Salton (to whom, as long-standing colleagues, I owe a great deal).

The work started well, and I continued to work in retrieval because, for everyone in NLP, the world turned cold in the later 1960s under the combined effects of the assaults on MT research and the fashion for Chomskyan linguistics. But it turned out to be much more complicated than expected to get term classes to do better than simple term matching in retrieving relevant documents. Though I could get benefits from classification for my first collection, when I tried others I could not get anywhere much.

Trying to understand why, I realized that term occurrence frequency, in the document file as a whole, not just cooccurrence frequency, was much more important than anyone had recognized.

For retrieval, you need to consider the discriminative as well as the descriptive value of a term, so you do not want to increase the matching power of a common term. As there are few relevant documents for a request among the many in the file, any term that occurs in a lot of documents is bound to retrieve more non-relevant than relevant ones. Allowing common terms to substitute for others compounds the problem.

This observation led to the proposal for term weighting, later called *idf* weighting, that turned out to be a big win: Terms should be weighted inversely by the number of documents in which they occur. The particular function, shown in Figure 5, is extremely, and pleasingly, simple but has been shown to be generally useful. Salton saw the value of *tf* weighting, that is, the value of term frequency within a single document, where the greater the frequency, the more important the term is for the

A statistical interpretation of term specificity and its application in retrieval, 1972

index terms variable matching value
terms in many documents good for recall
terms in few documents good for precision
weight terms by *inverse* document frequency

$$idf(t) = log(N/n)$$

Figure 5

document. With both ideas about weighting, researchers were beginning to get a better idea of the significance of a word's text statistics.

This was only within IR. Computational linguistics (CL) was focused on sentence grammar, artificial intelligence on world knowledge, and text interpretation from that point of view.

But in IR, the weighting idea had yet more to it. *Idf* weighting tries to use general term frequency data, for a whole collection, to predict specific query relevance. If you have some past relevance information, you can do better.

As with classification, it is tricky to get the formula right. I went through four successive versions, in an experiment/theory interaction with Stephen Robertson, at the beginning of a long and valued collaboration. But the outcome was very satisfying, not just for the performance results, but because Robertson's Probabilistic Model for retrieval relates the task to the text data in a convincing way. Figure 6 shows how effective this relevance weighting approach is, even for a difficult later data set using only document titles though rather rich requests. Even with only a few known relevant items to supply information about query terms in an iterative search, performance is much better than with *idf* weights alone, and when there are many known relevant to exploit, performance is strikingly better. The results illustrate the value of having training data for what would now be called a machine-learning strategy. Full relevance information can also be used to search retrospectively, to define a rational upper bound for performance.

This work in the 1970s looked very good, but there was a practical challenge in scaling experiments up to a realistic level, because testing needs reference data in the form of relevant documents for requests, and getting this for large collections is expensive. Operational services were not interested, partly because they believed in older conventional methods and partly because they had genuine other concerns like efficiency. The CL community was addressing the quite different goal of building natural language front ends to databases, which seemed a much more important project, addressing a task that needed proper syntactic and semantic interpretation. I thought it was an interestingly distinct challenge and joined in.

Natural language access to databases, converting a text question into, for example, SQL, looked like a tractable task at the time. But it turned out much less so than

S.E. Robertson and K. Sparck Jones
Relevance weighting of search terms, 1976

weight terms using relative frequency in
relevant and non-relevant documents

75 x 27 K chemical data:

	P at R = 30
idf	.11
few rel	.24
many rel	.44
upper bound	.69

[KSJ, 1979]

Figure 6

expected because it needs a large, rich domain model to bridge the gap between the user's free input and the particular data model, and also to trap the user's unwitting excursions outside the database boundary.

The difficulties of doing data access well also emphasized how narrow and limited database query, as normally implemented, actually is as a form of information access. It is much more reasonable to envisage a family of access modes to different types of information giving the user whatever can be found, in one way or another, in text, data, or even knowledge bases, as these happen to be available. A system would take the user's single input and apply different types of interpretation at different levels, within a single framework. It would treat a query deliberately as something to be unpacked from different points of view, rather than in a preferred mode, with backup defaults, as in LUNAR (Woods 1978).

Figure 7 illustrates the idea, very simplistically. Thus for the user's input, *Who supplies green parts?*, suppose we have an initial interpretation, say as a dependency tree. This can be subjected to further processing as a whole to construct a formal database query or taken simply as a base from which variant phrasal terms for text searching can be derived. Alternatively, if we consider the input text itself as its own, direct, representation, we can use this representation as a source for simple retrieval terms. Other versions of the input are drawn from different, perhaps successively deeper, representations as needed for particular tasks. Treating the text itself as a representation emphasizes the point that document retrieval is one proper task in its own right, among several, not a poor substitute for "true" question answering.

The idea that one can usefully work on the linguistic surface was reinforced in the 1980s by the experience of trying to build interfaces to more complex (expert) systems, for instance for advice.

The presumption was that we need to model the user's beliefs, goals, etc., to ensure an appropriate system response. It is not enough to take an input's "obvious" meaning. But in many cases, the system cannot get enough information about the user's knowledge, desires, or intentions to make reliable inferences about input motivations. So it may be more sensible to adopt a conservative system response strategy.

For example, suppose we have a travel inquiry system, with the user asking about trains, as in Figure 8. The system, seeking the reason behind the user's *Can I travel to Istanbul by train?*, might hypothesize that the user wants train travel because he or she

Shifting meaning representations, 1983

multiple levels of representation
    eg linguistic, logical
multiple task uses
    eg database query, retrieval term query

Who supplies green parts?
    give (agent [?], (object
        (be (object [parts] state [green])))
    = > for every V1/? (for every V2/part ..

    = >> green parts / parts that are green / ...

Figure 7

Tailoring output to the user: what does user modelling in generation mean?, 1991

interactive (advice etc) dialogue
model user's objective/subjective properties
to tailor system response?
respond to user's overt behaviour?

Can I travel to Istanbul by train?
Its slow and expensive.
Yes, there are three trains every day.

Figure 8

believes that trains are fast and cheap and would therefore, using its own information about what train travel is actually like, reply that it is slow and expensive. But the user may in fact have quite different (of many possible) reasons for wanting to travel by train, like wanting to see the scenery, and find the system's response distinctly inappropriate. The system would do much better simply by giving a straightforward response, perhaps adding some immediately pertinent amplifying detail.

The basic NLP tools for tasks like this, at the sentence and local discourse level, were improving all the time. But the suggestion that you don't need to dig very deep in language interpretation for useful task systems was indicative of a coming change.

In fact, there was an upheaval in the early 1990s like that of the early 1960s. Language and information processing (LIP) in the 1960s had had incipient machine text, and computers came on stream. In the 1990s we had enormously more machine power, and bulk text came on stream. We now have the whole Web as a language data source and arena for applications. But already in the early 1990s, more text and processor power stimulated the community to look again, as not since the 1960s for both practical and intellectual reasons, at statistical approaches to LIP. This was not just for resources, e.g., finding lexical associations or building probabilistic grammars, but for task systems, especially task systems responding to bulk text, like information extraction and summarizing.

Both of these are challenges to language processing, but especially summarizing, as for summarizing we need to understand large-scale discourse structure to do it, and there are different forms or aspects of such structure to investigate. For example, we can distinguish structure of a primarily linguistic type from communicative structure and from structure referring to the world. The same text can be characterized in different ways using these kinds of structure, as illustrated in Figure 9; and we can exploit each of these structure types for summarizing, since the way they are instantiated in the text can indicate more or less important content.

Thus considering the source text (even as abbreviated) shown in Figure 9, we can identify a linguistic structure of parallel description, saying something, in similar style, about biographies, about histories, and about novels. We can also see a communicative structure of the form: Say *X* to motivate act *Y*. And there is a further structure to the world being talked about, which characterizes books through their properties and uses. Each of these structures conveys information, through its form and associated content detail, about what is more or less important in the source text from the relevant point of view, and can therefore be used for summarizing. Thus the linguistic structure leads to a contrastive summary emphasizing key book-type features (LS), the

Discourse modelling for automatic summarising, 1995

discourse structure types
linguistic, communicative, world ...
role in marking important content
relative effects, merits for summarising

Biographies are the best books. They are about real things. They tell a true story. They are about particular people. History books ... true ... not about people. Novels .. not true ... about people. Give children biographies ... not novels ...

linguistic structure	: parallel description
communicative	: motivate action
world	: book types, uses

LS *Biographies are true and about people, histories true, novels untrue. Give children biographies.*
CS *Biographies are true and about individuals so give them to children*
WS *Biographies, histories and novels are both like and unlike. Biographies are good for children.*

Figure 9

communicative structure to a justification for action summary (CS), and the world structure to a simple descriptive summary (WS), without the presentational or functional character of the other two. The three summaries have something in common, because they all deal with what the text is about. They are also all different, but equally convincing as summaries, albeit from their different points of view. (Of course this illustration is a mere indicative sketch, where the reality would be far more complex.)

But the most obvious task to work on at the beginning of the 1990s was retrieval. Could the established research methods scale up from thousands of items to select relevant documents from millions? The TREC (Text REtrieval Conferences) evaluation program (Voorhees and Harman, in press) was designed to answer this question, but its scale and development, in a major, long-term activity under Donna Harman and Ellen Voorhees, have had a far wider impact on the whole of LIP.

I was delighted to see that the Probabilistic Model, in work led by Stephen Robertson, did just fine in TREC, combining *tf* with *idf*, as Salton had advocated, in a robust way and incorporating query expansion in feedback, as originally suggested by Rocchio in the 1960s (Rocchio 1966). Figure 10 shows how effective the statistical approach to IR can be (using results taken from a later paper). For two versions of the users' information needs, minimal and amplified, the performance gains are similar, though with richer starting requests they are, not suprisingly, larger. Performance for the rock bottom baseline, unweighted terms with a best-match strategy, gives only 1 relevant item in the top 10 retrieved. Using *tf* and *idf*, with a suitable elaboration to factor document length (*dl*) into the formula suggested by the theory, immediately gives a very substantial improvement, so half the top 10 retrieved are relevant. Feedback using only a small sample of known relevant documents gives a further gain when the queries are expanded with terms from these documents. Finally, it may be

K. Sparck Jones, S. Walker and S.E. Robertson
A probabilistic model of information retrieval: development and comparative experiments. Parts 1 and 2, 2000

probabilistic model, large scale tests
systematic comparisons with term weights

150 requests, 370 K documents, full text:

	precision at rank 10	
	10 terms	4 terms
unweighted terms	.11	.15
basic weighted	.52	.47
relevance weighted, expanded	.61	.51
assumed relevant	.57	.46

Figure 10

sufficient, with good starting requests, simply to *assume* that the best-matching documents in an initial pass are relevant, without invoking the user, to get a feedback gain.

These statistical techniques are very easy to apply, and the basic *tf * idf with dl* scheme was taken from our research for the first serious Web search engine, AltaVista, though 25 years after my *idf* paper, showing how long research can take for timely exploitation even in the rapidly moving IT world.

But the TREC program became more than just a regular document retrieval project. It has expanded over data types (e.g., to spoken documents) and into related tasks like question answering. This is important, because it has brought retrieval in from the cold to the NLP community, and encouraged a more generic view of it.

It is also, and more, important because it has shown how powerful the statistical approach to LIP is, and how widely it can be applied. Progress with statistical speech recognition and machine learning have helped in this, but retrieval has been vital because it engages with text content and has exported ideas about ways of handling this elsewhere.

Over the 1990s there has been a real growth in the use of statistics for LIP tasks. This has included some research on automatic lexical classification, i.e., on resource building, though ironically the main, widely used generic classification, WordNet, is manually built. But there has been far more emphasis on statistically determined word importance in text and also on statistically justified word sequences, leaving class relations implicit. Word importance, and the associated extractive view of text meaning, has turned out to be very valuable.

With rough tasks, like retrieval, you can do the whole job statistically, and well. With other tasks you can do a crude but useful job, wholly or largely statistically, for example by combining statistics with lightweight parsing in extractive summarizing or question answering. Using statistics for summarizing goes back to Luhn (1958). Figure 11 shows Luhn's "auto-abstract" for the 1958 conference paper with which I began. Concatenating source sentences that have been selected because they contain statistically important words does not make for readable abstracts. But the important fact about the example is that the mechanism for selecting sentences has chosen ones containing *thesaurus*, which does not occur in the paper title but is the focus of the text

IC51 INTERNATIONAL CONFERENCE ON SCIENTIFIC INFORMATION
AREA 5 PG 103
THE ANALOGY BETWEEN MECHANICAL TRANSLATION AND LIBRARY RETRIEVAL
MASTERMAN M CAMBRIDGE LANGUAGE RESEARCH UNIT CAMBRIDGE ENGLAND
NEEDHAM RM CAMBRIDGE LANGUAGE RESEARCH UNIT CAMBRIDGE ENGLAND
JONES ES CAMBRIDGE LANGUAGE RESEARCH UNIT CAMBRIDGE ENGLAND

AUTO ABSTRACT

NOW, IN THE PRESENT STATE OF RESEARCH THIS ANALOGY CAN ONLY BE DRAWN AT ALL PRECISELY BETWEEN ONE FORM OF LIBRARY RETRIEVAL PROCEDURE, AND ONE FORM OF MECHANICAL TRANSLATION PROCEDURE., THESE TWO ANALOGOUS PROCEDURES ARE THOSE, IN EACH FIELD, WHICH MAKE USE OF A THESAURUS.

WE PROPOSE, THEN, THAT A CONCEPTUALLY BASED, THESAURUS TYPE OF LANGUAGE CLASSIFICATION SHOULD BE USED FOR A COMPLETELY GENERALISED RETRIEVAL PROCEDURE, THIS CLASSIFICATION PROCEDURE BEING, BY ITS NATURE, INTERLINGUAL.

TRANSLATION SPECIALISTS, AND, IN PARTICUALR, LINGUISTS DENY EVEN THE POSSIBILITY OF THE ANALOGY BY MAINTAINING THAT ANY CLASSIFICATION OF LANGUAGE BASED ON A THESAURUS CAN, AT BEST, ONLY HOPE TO TRANSLATE SEMANTIC MEANING, WHEREAS LANGUAGE IS PRIMARILY A SYSTEM OF GRAMMAR AND SYNTAX., AND BOTH OF THESE ARE NOTORIOUSLY MONOLINGUAL.

THE OBJECT OF THIS PAPER IS TO REFUTE THIS CRITICISM BY SHOWING HOW A TYPE OF RETRIEVAL PROCEDURE, BASED ON A THESAURUS ALREADY BEING USED FOR THE EXPERIMENTAL TRANSLATION OF SEMANTIC MEANING, MIGHT ALSO BE EXTENDED SO AS TO TRANSLATE GRAMMAR AND SYNTAX.

Figure 11

argument. Modern techniques can combine statistical methods for identifying key content words with light sentence processing, e.g., to prune peripheral material, for more compact and coherent summaries.

Further, even with more sophisticated uses of NLP, for instance for question answering, searching bulk data means you need preliminary retrieval, which can be statistical, to find text passages likely to contain the answer, and may benefit from exploiting other statistical data, e.g., to identify related words or check proposed word relationships. In all of this, retrieval has supplied not only a general view of text, but specific techniques, notably *tf* * *idf*–style weighting.

Retrieval has also played a major role in research in the 1990s through its experience of system evaluation. This has not been just as an exporter of the often-misused notions of recall and precision, but through its emphasis on careful methodology and on evaluation geared to the task's function in its context. Thus retrieval is not about indexing per se, but about delivering relevant documents to users. Figure 12 sketches, in a very abbreviated form drawn from a fuller illustration, the kind of decompositional analysis required to design and conduct an evaluation of a task system operating in some context, from a particular point of view.

Thus if we imagine having a natural language interface to a house design system, we suppose that we want to discover whether this is effective as a training device for architecture students, and choose comparison with a menu system as the way of doing this. These decisions, and others, form part of the remit for the evaluation. They have to be fleshed out in the detailed design, which has to take account of environment factors, that is, of the values for environmental variables, like the difficulty of the planning problems set for students to tackle with the system, and of the settings for the system parameters, notably in this case the alternative interfaces being studied. The design also covers the choice of generic performance measures, e.g., output plan

K. Sparck Jones and J.R. Galliers
Evaluating Natural Language Processing Systems, 1996

decompositional approach
evaluation remit, design
environment variables, system parameters
performance criteria, test data and process

NL interface for house plan design:

remit:	training effectiveness, compare with menu
design:	environment - plan difficulty
	system - interaction modes
	criteria - plan quality ...

Figure 12

quality, and the particular ways of measuring this, along with the specification of the test data and of the procedure for carrying out the evaluation, e.g., choosing the planning problem sample, the students, etc.

As even this brief outline suggests, proper evaluation is a complex and challenging business. It implies, in particular, that we need to make a very rigorous "deconstructive" analysis of all the factors that affect the system being tested, as, for example, in the summarizing case sketched in Figure 13.

Here we have a particular purpose in summarizing, implying a specific function, namely, to alert, and a specific audience (i.e., readership), namely, the police; and in practice there are other purpose factors as well. Summarizing to serve this purpose has to take a whole range of input source document factors into account, like their subject domain and text form, in this case, weather reports. The purpose imposes constraints on output summaries, but there are still specific choices to be made about output factors like the language style and summary length, etc., which we suppose here leads to the production of very brief items in telegraphese. A breakdown of all the factors affecting a system is essential to guide evaluation.

Automatic summarising: factors and directions, 1999

input factors - subject type, form
purpose factors - audience, use
output factors - material, format
purpose + input constrain output

input	- weather forecasts
purpose	- weather alerts to police
output	- brief telegraphese

There will be long periods of heavy rain throughout the day in all areas.

$\Longrightarrow$ *warning : heavy rain everywhere*
prepare for accidents

Figure 13

External developments, in particular the huge growth of miscellaneous stuff on the Web, and the arrival of end users bypassing professional intermediaries, have encouraged the simple surface type of approach to LIP, as a general trend. But retrieval has been significant because it emphasizes the real status of text: Language meaning is behind text because it is also on the text surface. IR has also emphasized the fact that at some point with LIP tasks, maybe not locally but somewhere along the line, the human user has to interpret and assess text content. This is not because systems are deficient, but because natural language systems are for humans. For example, humans have to interpret and assess the answers to questions, even when these are "correct" in some obvious sense, in just the same way that they have to interpret and assess, albeit more elaborately and with more inference, a response text where the answer may be only approximate, qualified, or too deeply embedded for system unpacking.

There are many intellectual challenges in understanding what one is doing with language and language use under the statistical approach, compared with the natural and easy rationales one can give for grammars and parsing. For example, how does surface word behavior relate to forms of discourse structure? We also need to explain, in a principled way, what happens when we combine statistical and syntactic (or semantic or pragmatic) description and processing. Figure 14 shows just some of the questions to answer, for example, "How do data patterns relate to linguistic actions?" or "How do data units match linguistic units?", drawn from a much longer list.

We need to explain what is happening with statistics in generic language description and processing, for resources and operations that may apply across tasks. But we equally have to do it for task systems. The spread of statistics is currently making this very interesting for purely statistical systems (though we must also do it for hybrid strategies). One can apply Bayes' theorem, as the classical statistical tool, to anything. But even if this works very well in practice, you need to say what the grounding model for the task is. With a properly grounded formal model for the task that explains why the statistics work, you can hope to push further than with the super-abstract account, or the purely ad hoc apparatus, that statistics can easily supply.

This is very much an issue for the presently fashionable use of so-called Language Modeling. The Probabilistic Model for retrieval is grounded in the task, that is, it relates the term distribution data to the probability of document relevance. Language Modeling for speech transcription, which is where the approach came from, has a convincing grounding model in the idea of recovering a corrupted signal. But the

K. Sparck Jones, G. Gazdar and R. Needham (eds),
Computers, language and speech: formal theories and statistical data, 2000

issues -
relating data patterns to linguistic actions
combining data-derived rules for
integrated grammars
applying statistics to all description levels
matching data units and linguistic units
....
i.e. merging numeric with non-numeric information

Figure 14

Language modelling's generative model: is it rational?, 2004

task as recovering a generating source
speech recognition recovers the words
retrieval recovers the relevant document
translation recovers the other language form
summarising recovers the key concept
question answering recovers the answer

plausible task models ?

Ser o no ser, esa es la cuestión

Figure 15

recovery idea is much less plausible when taken as a justification for Language Modeling for summarizing, translation, or other tasks, under the interpretations shown in Figure 15. Is summarizing no more than recovering the original crisp few-liner from a mass of verbiage? More strikingly, is translating Shakespeare into Spanish just recovering, from the defective English, the Spanish in which he originally wrote?

As this current, active research implies, there is a lot of challenging work to do. I am very happy to have been in at the beginning of automated LIP, I am happy to be still in it now, and I am happy to have plenty more to look forward to.

In conclusion, I would first like to thank all my research students and assistants, from whom I have learned so much.

Then, referring again to the author names on my first paper, I want especially to thank Margaret Masterman who employed me at the start when the only qualification for research in LIP I had was a year reading philosophy (though this was a good qualification in fact). But I want most of all to thank my late husband Roger Needham, not only because we worked and published together at particular times, but because I could always talk to him about my research, and he always encouraged me.

Thank you again.

References

Items cited in the figures (in citation order)

Masterman, Margaret, Roger M. Needham, and Karen Spärck Jones. 1959. The analogy between mechanical translation and library retrieval. In *Proceedings of the International Conference on Scientific Information* (1958), National Academy of Sciences–National Research Council, Washington, DC, Vol. 2, pages 917–935.

Spärck Jones, Karen. 1964. *Synonymy and Semantic Classification*. Ph.D. thesis, University of Cambridge: Report ML 170, Cambridge Language Research Unit. (Edinburgh University Press, Edinburgh, 1986.)

Spärck Jones, Karen. 1967. A small semantic classification experiment using cooccurrence data. Report ML 196, Cambridge Language Research Unit, Cambridge.

Spärck Jones, Karen. 1967. Notes on semantic discourse structure. Report SP-2714, System Development Corporation, Santa Monica, CA.

Spärck Jones, Karen. 1972. A statistical interpretation of term specificity and its application in retrieval. *Journal of Documentation*, 28:11–21.

Robertson, Stephen E. and Karen Spärck Jones. 1976. Relevance weighting of search

terms. *Journal of the American Society for Information Science*, 27:129–146.
Spärck Jones, Karen. 1979. Search term relevance weighting given little relevance information. *Journal of Documentation*, 35:30–48.
Spärck Jones, Karen. 1983. Shifting meaning representations. In *IJCAI-83, Proceedings of the Eighth International Joint Conference on Artificial Intelligence,* Karlsruhe, Germany, pages 621–623.
Spärck Jones, Karen. 1991. Tailoring output to the user: What does user modelling in generation mean? In C. L. Paris, W. R. Swartout, and W. C. Mann, editors, *Natural Language Generation in Artificial Intelligence and Computational Linguistics*. Kluwer, Dordrecht, pages 201–225.
Spärck Jones, Karen. 1995. Discourse modelling for automatic summarising. In Eva Hajicova, Miroslav Cervenka, Oldrich Leska, and Petr Sgall, editors, *Travaux du Cercle Linguistique de Prague* (Prague Linguistic Circle Papers), New Series, Volume 1. John Benjamins, Amsterdam, pages 201–227.
Spärck Jones, Karen and Julia R. Galliers. 1996. *Evaluating Natural Language Processing Systems*. Lecture Notes in Artificial Intelligence 1083. Springer-Verlag, Berlin.
Spärck Jones, Karen. 1999. Automatic summarising: Factors and directions. In I. Mani and M. T. Maybury, editors, *Advances in Automatic Text Summarisation,* MIT Press, Cambridge, MA, pages 112.
Spärck Jones, Karen, Gerald Gazdar, and Roger M. Needham, editors. 2000. Computers, language and speech: Formal theories and statistical data. *Philosophical Transactions of the Royal Society of London,* Series A, 358(1769): 1225–1431.
Spärck Jones, Karen, Stephen Walker, and Stephen E. Robertson. 2000. A probabilistic model of information retrieval: Development and comparative experiments (Parts 1 and 2). *Information Processing and Management,* 36(6):779–840.
Spärck Jones, Karen. 2004. Language modelling's generative model: Is it rational? Working paper, Computer Laboratory, University of Cambridge.

Other items cited in the paper

Luhn, Hans Peter. 1957. A statistical approach to mechanised literature searching. *IBM Journal of Research and Development,* 1(4):309–317.
Luhn, Hans Peter. 1958. An experiment in auto-abstracting. Auto-abstracts of Area 5 conference papers. International Conference on Scientific Information, Washington, DC, November 16–21, 1958. IBM Research Center, Yorktown Heights, NY. (Reprinted in C. K. Schultz, editor, *H. P. Luhn: Pioneer of Information Science,* Spartan Books, New York, 1968, pages 145–163.)
Rocchio, Joseph J. 1966. *Document Retrieval Systems—Optimization and Evaluation.* Report ISR-10, Computation Laboratory, Harvard University.
Voorhees, Ellen M. and Donna K. Harman, editors. In press. *TREC: An Experiment and Evaluation in Information Retrieval.* MIT Press, Cambridge, MA.
Woods, William A. 1978. Semantics and quantification in natural language question answering. In M. Yovits, editor, *Advances in Computers,* Vol. 17. Academic Press, New York, pages 2–64.