LINGUISTICS AND INFORMATION SCIENCE: A POSTSCRIPT

Karen Sparck Jones
Cambridge University

and

Martin Kay
XEROX Palo Alto Research Center

The object of *Linguistics and Information Science* (Sparck Jones and Kay, 1973) was to show how far the suppposedly natural connection between linguistics and information science existed in practice. We surveyed linguistic theory and computational linguistics to identify approaches potentially applicable to information science, and to information, i.e., document, retrieval in particular; and we investigated the linguistic operations of automatic document retrieval to establish their linguistic sophistication and the extent to which linguistic theories were being. or could be, applied. We also looked for evidence of feedback from automatic information retrieval to linguistics. Our general conclusion was that there was very little actual connection between linguistics and information retrieval. Linguists were preoccupied by concerns rather remote from any practical activity like information retrieval, for example the properties of linguistic theories, and had failed to provide tools of potential utility to retrieval workers. At the same time, in both practice and research in information retrieval, needs which might be met by linguistic theory were not properly specified. In general, the linguistic procedures of automatic information retrieval were found to be very simple, and it was not obvious how useful refined linguistic tools would be, either as aids to automation, or as devices for improving retrieval performance.

In this note, we shall briefly consider the major developments in linguistics and information retrieval that have taken place since we completed the manuscript of our book in mid 1971, to see whether linguistics and information science are, or could be, more closely linked now than they were then. The note is not intended to be a detailed survey, but rather a set of comments within the general framework of *Linguistics and Information Science*. Our remarks are therefore independent of the other papers in this volume, and the reader is referred to these papers for more detailed treatments of individual topics, from different points of view.

Linguistics

Theoretical Linguistics. In 1971, transformational grammar, in one version or another, was clearly the dominant linguistic theory in North America and was gaining adherents throughout the world at a rapid rate. Even in countries like England, Denmark, and

Czechoslovakia, with vigorous linguistic traditions of their own, it appealed to many as the paradigm most likely to shed light on outstanding linguistic problems. However, already by the time our book was published, the term transformational grammar could no longer be used to refer to a single coherent body of doctrine. In *Aspects of the Theory of Syntax*, Chomsky, (1965), proposed the theory according to which sentences had underlying structures generated by a context-free base component. The sentences themselves were obtained from these by the application of transformational rules which did not, however, have any effect on meaning. Optional tranformations made it possible to derive sets of two or more sentences from the same underlying structure, but, in these cases, it was claimed that all the sentences in the set would have the same semantic interpretation. It was an appealingly simple view and constituted a strong claim about the nature of human language. But it proved impossible to uphold this claim. It soon appeared that certain aspects of semantic interpretation, notably those concerning the use of quantifiers, depended in crucial ways on the surface forms of sentences. So, for example, "Every command is represented by a single code" is to be interpreted quite differently from "A single code represents every command". "He sent his daughter her allowance" has a different range of meanings from "He sent her allowance to his daughter".

Attempts to adjust the theory to account for facts such as these took a great variety of forms. The proponents of *generative semantics* pursued the view that deep structures should be more "abstract", that is, more remote from the sentences they underlie and that a more complex transformational apparatus should relate them to surface forms. According to this view, no separate semantic component is required because the deep structure *is* the semantic interpretation. This was typically combined with the proposal that the theory specify a set of *transderivational constraints* restricting the sequences of transformational rules that could apply to a given deep structure. The introduction of such heavy machinery could be justified, in the face of the requirement for explanatory adequacy, only if these contraints were part of the overall linguistic theory and not a separate part of the grammar of each language.

Another view was that the requirement that transformations should play no semantic role, and that the meaning of a sentence should be derivable entirely from its deep structure, should be weakened or abandoned. Jackendoff (1972), for example, proposed a scheme according to which the meaning of a sentence would be derived from a number of its representations, including the deep and surface structures, and also other structures that arise in the course of the transformational process.

However, it rapidly became clear that there was a great deal more than quantifiers and related logical problems to embarrass the theoreticians. They became increasingly impressed by the fact, never doubted by their colleagues in Europe, that the notion of semantic equivalence cannot be identified with that of logical equivalence. American linguists became increasingly interested in what has been called functionalism, broadly, the ways in which various kinds of utterance suit themselves to achieving the various goals that a speaker might have. This was stimulated, in part, by the notion of speech acts put forward by Austin (1965) and Searle (1969). It is observed that the difference between the sentence "I will be in the office tomorrow at noon" and "I promise to be in the office tomorrrow at noon" comes not from any difference in the truth values of the propositions they represent, but from the nature of the commitment in which they engage me, the speaker. Predictably, the response of the generative semanticists was to decorate the already overgrown trees they proposed as the deep representations of sentences with a new layer of structure to accomodate performative verbs. "John ran" came to have a structure more like the one that would previously have been ascribed to "I assert that John ran", in which the performative "assert' declares the kind of commitment that the speaker has to what

follows.

A related notion is that of presupposition. The sentence "I like your new car" is true if I like it and false if I do not. But what if you do not have a new car? In this case, neither the sentence "I like your new car", nor its negation "I do not like your new car" is true. The purely contingent fact that you do not have a new car cannot render the sentence meaningless. The trouble is that both sentences imply that you have a new car and, because this implication is false, the phrase, "your new car" fails to refer properly. If a sentence and its negation both imply some proposition, they are said to presuppose that proposition. The notion of presupposition impacts the interpretation of natural sentences in various and often subtle ways. The sentence "Brutus killed Caesar" and "It was Brutus that killed Caesar" have the same truth value--each is true if and only if the other is--but the second presupposes that someone killed Caesar whereas the first does not.

Speech acts and presupposition belong to a class of essentially pragmatic phenomena whose study cannot be confined to the limits of single sentences. They are phenomena that cannot be summarily excluded from the study of linguistics for, just as it proved impossible to conduct a deep investigation of syntax without regard to semantics, so now it proves impossible to investigate semantics satisfactorily without regard to pragmatics. As a result, linguists find themselves committed to a view of language in which the status of individual sentences is greatly reduced.

These problems are still very poorly understood and certainly no theory with the formal elegance of transformational grammar has been proposed to accommodate them. A pessimistic appraisal of the resulting situation is that American linguistics is in a state of complete disarray with no common body of doctrine to unite even small groups of theoreticians. As viewed from London or Prague, the situation might appear more encouraging if only because, from these vantage points, transformational linguists have at least demonstrated the maturity to face what are, after all, quite old problems. On the other hand, the considerably longer time that Halliday, Firbas, Sgall, and their European colleagues have spent with these problems has provided little in the way of solutions.

Computational Linguistics. Sparck Jones and Kay (1973), adopts the characterization of computational linguistics, due to Hays, (1967), as "those linguistic activities in which the computer plays a central role". A better characterization, especially in view of more recent work, might be as linguistics in which computation provides a major source of inspiration. The breakdown of the transformational paradigm and the need which many more linguists now feel to examine linguistic phenomena in a wider context have done much to undermine the distinction between competence and performance. Consequently, there is renewed interest in studying the strategies that people employ in producing and understanding utterances as well as abstract constraints on the forms they can take. It is not surprising that this line of attack has never been pursued far in the past because the vocabulary and metaphors necessary to investigate complex abstract processes were simply not available until they were provided by computer science. The term "computational linguistics" can therefore be properly applied to linguistic activities that do not involve actual machines at all. To the extent that they make use of notions of variable binding, control structure, process scheduling, and the like, they are computational.

Starting from an original idea of Thorne, Bratley and Dewar (1968), further developed by Bobrow and Fraser (1969), Woods (1970), developed a parsing scheme based on what he calls an Augmented Transition Network Grammar (ATN) (see also Sparck Jones and Kay, 1973, pp. 100-101). This parser was incorporated into at least two question-answering programs, one for the U.S. Defense Documentation Center and the other, for use on

geological information collected on the moon, for NASA (Woods et al., 1972). These systems attracted much attention because of the unprecedentedly wide coverage of their dictionaries and grammars and also because of the overall smoothness and efficiency with which they worked. The English grammar that these systems incorporated was written largely by Kaplan who also wrote the grammar of the MIND system (see Kay, 1967). Based on similarities that he perceived in these apparently very different systems, Kaplan (1973), designed what he called the General Syntactic Processor which generalizes and greatly simplifies the two preceding techniques. Furthermore, it makes a clear distinction between the grammatical rules approximately modeling a speaker or hearer's competance and a set of scheduling rules which, by determining which of the possible courses of action open to the processor at any given moment will be followed, enshrine the processing strategies.

At the same time, a new syntactic formalism, called relational grammar was being developed by Postal and Perlmutter (forthcoming) which Lakoff and Thompson (1975), saw as having a strong family resemblance with the grammars of Woods and Kaplan. It thus appears that what began as an engineering device to meet very practical goals has, through a series of successive refinements, become a theoretical contribution to linguistics. The interest of this particular contribution is that it provides a basis upon which to make predictions about details of human linguistic behaviour which are susceptible of experimental verification (see Kaplan, 1972).

At the beginning of the period covered by this postscript, Winograd (1972), completed his SHRDLU system, a computer program which manipulates children's building blocks (as depicted on a two-dimensional display) in response to directions given to it in English. The interest of the program lies mainly in its ability to make simple but natural inferences about what it is told to do. If it is told to put the red block on the green one, but the green one has a yellow one on top of it already, then it infers that it must remove this first. Sometimes it must make quite complex preparations before the final result it has been directed to achieve can be reached. Asked why it made this or that move, it will respond with an acceptable explanation. Thus, if one asks "Why did you move the yellow block?" it will say something like, "Because I wanted to put the red block there." The system will respond correctly to commands to build relatively complex structures, like towers, out of the blocks and will accept definitions of hitherto unknown words denoting new structures to be built.

Also during this period, a very considerable amount of money has been spent by the Advanced Projects Research Agency of the US Department of Defense on research directed towards speech understanding (Newell *et al.*, 1973). Speech understanding is to be clearly distinguished from speech recognition in that it aims to go beyond the recognition of isolated spoken words and to respond in some potentially useful way to connected speech. This work, conducted at a number of centres in the United States, led to a considerably increased understanding of how phonetic and phonological data might be processed and how this might be integrated with work on semantics and syntax already under way. In particular, it led to the development of more flexible methods of managing the interaction between the various components of a large computer system. Very early in this work, it became clear that the established view that a linguistic processor could consist of a series of components, each taking as input the output of the one before, was no longer viable. Recognition of phonetic units in the acoustic signal required not only the data in that signal, but also predictions about what later parts of the signal might contain based on a syntactic and semantic analysis of what had already been heard. It was necessary to devise schemes whereby the components could work essentially in parallel, each being prepared to accept and deliver partial results, hypotheses, and predictions, even at a very early stage in the analysis.

The systems we have been discussing had the common property that they could process successfully only carefully constructed sentences about the very narrow subject matter for which they were designed. They were designed around what has come to be known as a *microworld*. The extent of their contribution to linguistics, computational or general, is therefore very much an open question. Such attempts as there have been to apply the methods of linguistics to larger universes of discourse or to larger subsets of the language have rested on much more modest goals in linguistic analysis. The extreme position is represented by a program called PARRY (Colby 1973) which plays the part of a paranoid patient in a dialog with a person who is expected to play the role of a psychiatrist. In this, as in much other psychiatry, the object is not to cure the patient, but to prolong the dialog. The program is accounted successful to the extent that it sustains the illusion of a dialog with a real patient. What is interesting about the program is that it rests on none of the advanced theories and complex techniques that are the basis of the other programs we have discussed. Yet it is the only one that will perform successfully in the hands of a naive and untrained user. This success is, of course, largely attributable to the carefully chosen scenario within which the program operates. The well-formedness conditions on a conversation between a psychiatrist and a sufficiently sick patient are presumably minimal. It is however interesting that the illusion of continuity can be maintained at all, and it at least suggests that a considerable amount may yet be achieved by relatively simple techniques provided only that they take into account whole texts or whole discourses rather than single sentences.

But, while recognizing the limitations of the approaches described, it can be said that there has been progress in computational linguistics since 1971; in particular, the systems implemented, though small, have been quite solid. Computational linguistics thus appears of more potential relevance to information retrieval than it did then.

Information Retrieval

In information retrieval, the most striking development since 1971 has been the growth of on-line search systems for very large data bases. The automation of library operations in general has continued, in both small and large libraries, a notable development being the growth of library networks, for example involving communal cataloguing, as in the Ohio College Library Center. But the development of on-line search systems is more interesting in the present context, since it depends on linguistically interesting procedures of indexing and searching. It is therefore all the more disappointing, from our point of view, that these systems do not generally involve sophisticated automatic linguistic operations. Their effectiveness derives in large part from their computational power, since they can scan extremely large files for items satisfying complicated specifications, in a manner which is quite impossible for the human library user. As Lancaster and Fayen's 1973 survey shows, these systems may offer a whole range of search keys, and depend on a variety of linguistic indexing modes, including on the one hand manual indexing using a controlled thesaurus, as in Medline, and on the other the exremely simple form of automatic indexing represented by the provision of unprocessed title and abstract texts.

Since these systems to a considerable extent simply provide mechanical support for the human searcher, rather than an automatic substitute for him, indexing can be seen as *request* rather than *document* oriented. This is particularly obvious when title or abstract texts are searched. Linguistic entities like word classes or phrases are generated for individual requests, and corresponding very fine partitions of the document set are established on searching rather than when documents are filed. While any request above the single word incorporates an element of a posteriori indexing, current automatic search systems may permit extremely complex specifications, including, for instance, very variable

truncation options, and so emphasize indexing as a request rather than a document-oriented process. At the same time, the fact that searching in on-line systems is interactive means that a whole range of information about words and documents is exploited, but not necessarily in a systematic or coherent manner. These systems are hopefully hospitable to a great variety of needs, and allow great flexibility in searching. They are, in consequence, extremely difficult to characterize in terms of language-using processes and to evaluate. The variable human elements involved mean that it is not at all easy to establish the relative value of different types of linguistic information or different linguistic procedures.

Perhaps the only strictly innovative feature of these systems is the provision of statistical information, e.g., about the collection frequency of index terms as a guide to their effective, as opposed to notional, discriminating power. Some tentative steps have been taken in the use of statistically based weighting schemes. Statistical information is, perhaps, of marginal linguistic interest, being essentially descriptive of the sublanguage represented by a document collection, rather than systematic. But it is one of the few distinctive contributions of the computer based approaches to indexing studied in the nineteen sixties. Statistical weighting is also an active research area. In the absence of richer, automatically obtained information, getting some mileage out of such easily obtained data is an attractive option, particularly since such experiments have been carried out, for instance by Salton (Salton and Yang, 1973) and Sparck Jones (1972), suggest that term occurrence information can be sucessfully exploited to improve retrieval performance. In *A Theory of Indexing*, Salton (1975), analyses the retrieval value of term frequencies in a systematic way, and argues that terms with frequencies in different ranges should be handled in different ways to modify their matching behaviour. Thus, if medium term frequencies are most useful for both recall and precision, low frequency terms may be grouped in classes to increase their effective frequency, and high frequency terms combined as phrases to reduce theirs. Roberts and Sparck Jones (1976), have suggested that retrieval performance may be further improved if weighting is based on information about relevant document distribution of terms, and report successful experiments along these lines.

Statistical term classification, on the other hand, has not been proven effective, and research on this and other topics of interest in the nineteen sixties seems to have declined. Work continues on document clustering, including that based on linguistically eccentric keys like citations. But, in general, the situation in information retrieval, and particularly in research, is very different from that in 1970. On the other hand, technological advances, like the provision of on-line searching appear to have removed the need for autonomous automatic document and request processing. Manual indexing and the use of manually constructed thesauri have also been maintained, partly through organizational inertia, partly because their cost can be spread over a great many information products, partly through continued faith in their merits, and partly because sufficiently comprehensive and rigourous experiments demonstrating their ineffectiveness, or showing that less exigent indexing techniques are superior, have not been carried out. An example of a continuing commitment to manual indexing, not merely because automatic substitutes have not yet been provided but because the human indexer is believed to be necessarily superior to a computer, is the British Library's PRECIS system (Austin, in this volume). It has also been recognized that retrieval system use and performance are affected, or determined, by other factors than the core linguistic ones of primary concern in the nineteen sixties. Irrespective of whether specific linguistic information can be obtained automatically as opposed to manually, its utility may be small. Librarians have doubtless known this all along, but in automatic retrieval systems, 'sociological' factors have only recently been given their due.

On the other hand, it must be accepted that retrieval research has not altogether produced

the goods. Sparck Jones 1974 survey in *Automatic Indexing* 1974 showed that comparatively few reasonably scaled and rigorous experiments have been carried out which individually and collectively provide comprehensive and systematic data on the relative merits of automatic and manual indexing, or linguistically complex and linguistically simple indexing. It is even true that adequate tests of complex and simple manual indexing are lacking. However, research workers can defend themselves by pointing to a manifest reluctance on the part of system operators to consider research problems or take much notice of apparently relevant results; and by noting the change in experimental requirements accompanying operational system growth. In the nineteen sixties, experiments on a few hundred documents might be deemed operationally relevant. They would not be acceptable now, and, while it is evident that larger tests are needed because scale effects in retrieval are not well understood, this makes retrieval research more expensive and time consuming. The design of experiments relevant to on-line searching also presents many problems, and it is not at all obvious how sociological factors in retrieval systems should be studied, and their significance determined.

Referring to the more detailed topics of *Linguistics and Information Science*, specific remarks about the linguistic elements in retrieval systems can be made as follows. Automatic syntactic analysis for indexing is currently virtually non-existent. Work on the full syntactic processing of particular kinds of data for particular purposes is being undertaken by, for instance, Sager (in this volume) which, if successful, might have some bearing on document indexing in general. Partial parsing to identify words or word strings as candidate indexing terms is being applied on a large scale by Klingbiel (1973a,b), and at the lowest level minimal syntactic information may be exploited for the production of printed indexes. Syntactic information may also be tacitly exploited via request formulations involving combinations of words in a specified order or in specified proximity. But, for a linguistic point of view, such work is small beer. The question as to how far syntactic information can be of general utility in document retrieval indeed remains unanswered, though for printed subject indexes, for example, it appears of value. Work on automatic syntactic analysis for information retrieval must thus depend on further analysis of the kind of syntactic information that is really needed for retrieval, and how it should be exploited.

Under the semantic head, the non-trivial selection of words from longer texts is rarely undertaken, all the non-function words of titles or abstracts being taken as index keys or leads to keys. (Specialized applications like legal or patent text searching are exceptions.) Automatic vocabulary formation and control has, in recent years, been approached chiefly through weighting, as described above. Clearly, some selection is achieved if terms are zero-weighted, and, more generally, control is achieved through the different values determined by weighting. It must be emphasized that vocabularies characterized by weighting techniques are sensitive to changes in the composition of a document collection in the way that a priori vocabularies are not; a term may be a good discriminator at one stage in the life of a collection, but a bad one at another. The automatic identification of term relations is not being very actively pursued. As with syntax, the apparent failure of plausible approaches to the automatic identification and use of semantic relations have incidentally cast doubt on the retrieval value of such relations, however identified. As noted, many manual thesauri continue to be used, but the real and relative values of the types of information they may contain have not been determined in detail, and experiments to obtain such information automatically therefore seem rather pointless.

Conclusion

In conclusion, can we say that the relation of linguistics and information retrieval in 1976 differs from that of 1971? The foregoing does not show that there is greater actual connection between the two now than there was then. However, the relative positions have changed. Linguistics, particularly computational linguistics, seems more potentially applicable to information retrieval than linguistics did in 1971. On the other hand, in information retrieval there appears to have been a retreat from linguistic sophistication, and there is hence less interest in automatic language processing techniques of the kind currently being studied by computational linguists.

We nevertheless do not feel that linguistics and information retrieval can have no connection, though we see the difficulty of making the connection more clearly than we did in 1971. Essentially, the concerns of information retrieval and current linguistics are of a different scale. Information retrieval is necessarily concerned with the larger features of the masses of information represented by document collections, and with gross characterizations of individual documents. Linguistics is typically concerned with a refined characterization of small universes and units of discourse. The techniques required to approach linguistic information on these different levels seem to be very different. At the same time, the scale distinction between the concerns of linguistics and information science are not consistent over different aspects of language processing; for instance, linguists may require a very large grammar to provide a sufficiently precise characterization of individual sentences, where information scientists might get by, for gross characterization, with a small one. Thus, at virtually every linguistic point, there is some substantial difference in scale.

However, we persist in our optimism and foresee greater possibilities for collaboration in the future. We take heart particularly from two facts: first, linguists are turning their attention more and more to larger units of discourse than the sentence, and second, on-line retrieval systems are likely to involve retrievable units smaller than traditional documents. We believe that the relevance of these fields to one another will become more apparant as the size of the text units they deal with becomes more commensurable.

References

Austin, J. L. *How to Do Things with Words.* London, Oxford University Press, 1965.

Bobrow, D. G., and Fraser, B. "An Automated State Transition Network Analysis Procedure". In Walker, D. E., and Norton, L. G., eds. *Proceedings of the International Joint Conference On Artificial Intelligence.* The MITRE Corporation, Bedford Mass., 1969.

Chomsky, N. *Aspects of the Theory of Syntax.* Cambridge, Mass.: MIT Press, 1965.

Colby, K. M. "Simulations of Belief Systems". In Colby and Schank, eds., *Computer Models of Thought and Language,* San Francisco: Freeman, 1973.

Hays, D. G. *Introduction to Computational Linguistics.* New York, American Elsevier, 1967.

Jackendoff, R. S. *Semantic Interpretation in Generative Grammar*, Cambridge, Mass.: MIT Press, 1972.

Kaplan, R. M. "Augmented Transition Networks as Psychological Models of Sentence Comprehension", *Artificial Intelligence*, 1972, 3, 77-100.

Kaplan, R. M., "A General Syntactic Processor". In Ed. Rustin, ed., *Natural Language Processing*, New York. Algorithmics Press, 1973.

Kay, M. *Experiments with a Powerful Parser.* RM-5452-PR, The RAND Corporation, Santa Monica, 1967.

Klingbiel, P. H. "Machine-Aided Indexing for Technical Literature". *Information Storage and Retrieval*, 9, 1973, 79-84.

Klingbiel, P. H., "A Technique for Machine-Aided Indexing", *Information Storage and Retrieval*, 1973, 9, 477-494.

Lakoff, G., and Thompson, H. "Introducing Cognitive Grammar". *Proceedings of the First Annual Meeting of the Berkeley Linguisitcs Society.* University of California, Berkeley, 1975.

Lancaster, F. W. and Fayen, E. G. *Information Retrieval On-line.* New York. Wiley, 1973.

Robertson, S. E. and Sparck Jones, K. "Relevance Weighting of Search Terms". *Journal of the ASIS*, 1976.

Salton, G. and Yang, C. S. "On the specification of Term Values in Automatic Indexing". Journal of Documentation, 1973, 29, 351-372.

Salton, G. *A Theory of Indexing.* Regional Conference Series in Applied Mathematics, No. 18, Society for Industrial and Applied Mathematics, Philadelphia, 1975.

Searle, J. R. *Speech Acts.* Cambridge, Cambridge University Press, 1969.

Sparck Jones, K. and Kay, M. *Linguistics and Information Science.* New York: Academic Press, 1973.

Sparck Jones, K. *Automatic Indexing 1974.* Computer Laboratory, University of Cambridge, 1974.

Thorne, J., Bratley, P., and Dewar H., "The Syntactic Analysis of English by Machine". In Michie, D., ed., *Machine Intelligence 3.* New York, American Elsevier, 1968.

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

Woods, W. A. "Transition Network Grammars for Natural Language Analysis", *Communications of the ACM*, 1970, 13, 591-606.

Woods, W. A., Kaplan, R. M., and Nash-Webber, B. *The Lunar Sciences Natural Language System.* Bolt, Beranek and Newman Inc., Cambridge, Mass., 1972.