THE ORGANIZATION AND CONDUCT OF THE WORKSHOP

Donald E. Walker
SRI International

The Goal

The goal of the FID/LD Workshop on Linguistics and Information Science was to contribute to the development of a comprehensive plan that can be used to guide research activities for more effective use of linguistics in information science. To appreciate our progress toward that goal during the three days at Biskops-Arnö in May of 1976, it is necessary to consider in some detail the stimulus that prompted the establishment of the Workshop, the composition of the group that participated, the common information we shared in the form of the materials distributed in advance to the participants, and the similarities and differences in our perceptions of the needs for linguistics in information science and of the resources that are available to work with. This chapter is organized so that these items of background data are presented first, followed by a description of the conduct of the Workshop itself.

The Starting Point

The point of departure for the Workshop was the survey of *Linguistics and Information Science* prepared by Karen Sparck Jones and Martin Kay (1973). It provided a baseline for considering the relationship between those two disciplines -- and related areas of inquiry, as of 1970. Thus, distributing the book to the participants in advance assured a common frame of reference both for the people who prepared perspective papers and for the Workshop participants themselves during our discussions. Consequently, a brief review of its objectives, its contents, and its conclusions seems in order.

Sparck Jones and Kay interpret *information science* as concerned broadly with "the storage, retrieval, and transmission of information of any kind" (p.2), but identify their interest more narrowly with *information retrieval.* Specifically, they deal with "the problems that arise in characterizing the content of documents and information requests in such a way that the characterizations can be used in an automatic process which can assess the relevance of the document to the request" (p.2). Key elements in this focus are the emphases on the *content* of documents and of information requests, on the *relevance* of a document to a particular request, and on the *automatic* nature of the procedure that could perform the coordination of document and request.

Linguistics for Sparck Jones and Kay is "the science that attempts to explain how language works" (p.3). Of particular concern, in relation to information science, are the ways in which linguistics can be expected to contribute to the naturalness of communication with a documentation system; to the reduction of ambiguity; to the precision of the description of a document, of the request for it, and, consequently, of the correspondence between these two elements; to the establishment of an adequate semantic structure; and to the separation of the language to communicate with the system from the data stored in it.

The thorough review of the literature in the body of the book demonstrates the minimal realization of these expectations by 1970. Following discussions of current information retrieval systems and relevant developments within linguistics at that time, the authors present a careful analysis of the way that language figures in documentation. They then show how work in syntax and semantics has been applied in information retrieval and where it might also be applied. Sparck Jones and Kay admit to being puzzled somewhat by the failure of systematic experiments to demonstrate that linguistic techniques have any real value in the contexts to which they have been applied.

In the final chapter of *Linguistics and Information Science*, Sparck Jones and Kay review their findings to determine why there has been such difficulty in applying linguistic techniques to retrieval objectives. Several points were made in their introduction that anticipated these negative conclusions about the contributions linguistics had made to information science. It certainly is the case that information scientists usually are concerned with large collections of document texts, while few linguists have dealt with units larger than sentences. Furthermore, information scientists wishing to demonstrate the effectiveness of their procedures deal with precision-recall ratios, while linguists trying to account for linguistic data try to determine what kinds of constraints are involved in the production of an utterance. Computational linguists, who use computers in the analysis of linguistic problems, could be expected to provide tools for information science. However, the most relevant work they have done is in the area of fact retrieval and question answering systems. Rather than responding to a request with a document or set of documents, the goal of such systems is to provide a particular fact or a precise answer. Since the specific items required are unlikely to be stored in the data base in precisely the form requested, inferences often have to be made. In view of the relative immaturity of work on computer-based theorem proving and deduction, much remains to be done for work on these problems to be useful.

On the basis of their analysis, the easiest course for Sparck Jones and Kay would have been simply to accept the conclusion that general linguistic theories are not required for document retrieval. However, they suggest that it would be more productive to recognize both that linguistic theories are far from adequate and that document retrieval systems are not well understood. Accordingly, the course they promote is to find out more about both areas. This same objective guided the FID/LD Workshop in its deliberations.

The Group

The participants in the Workshop were selected to reflect the variety of backgrounds, interests, and responsibilities relevant to research on linguistics in information science. Thus, we invited people from theoretical, computational, and quantitative linguistics, from information science, library science, and documentation, from terminology and translation, and from computer science and artificial intelligence. Of course, no discipline is one dimensional, so this listing does not adequately portray the diversity of interests that were represented. In addition, the selection was influenced by the international nature of the organizing body. Not only are there FID/LD Committee members currently representing eleven countries (nine of them were present at the Workshop), but research in linguistics and information science and applications of linguistics to information science take place in many countries throughout the world. Competence is distributed without respect for national boundaries, and the potential utility of results in these areas of endeavor is recognized throughout the more industrialized nations of the world.

The Organization and Conduct of the Workshop

There were thirty participants at the Workshop, from thirteen countries: Belgium, Denmark, France, the German Federal Republic, Hungary, Israel, Italy, Luxembourg, the Netherlands, Norway, Sweden, the United Kingdom, and the United States. These differences in nationality certainly provided a leavening to our deliberations, but their effects were minimal in comparison with the effects of differences in discipline and, perhaps most important, of the variations in the characteristics of the professional activities people were engaged in.

It is not possible to provide a matrix that will identify the participants uniquely in relation to their professional activities. Rather, by identifying some of the dimensions along which their responsibilities contrasted, I hope to indicate the richness and complexity of the group. The variety of disciplines has been identified already; within them there were differences between theoreticians and practitioners. In relation to design, the contrast would be between formal elegance and practical efficiency. A pair of relevant system-oriented terms is technology-driven versus demand-driven; these terms contrast orientations motivated by the emerging technology with those responding to requirements expressed by the user.

The nature and sizes of the data bases with which people dealt differed. Some participants were concerned with large document collections, some with experimental document systems, some with narrowly defined microworlds not even in documentary form, some with data bases they generated as required to provide examples and counterexamples relevant for specific theoretical issues. These differences reflected preoccupations variously with the global requirements of a comprehensive national library, with access to special document collections, with the selection of a set of documents that could be expected to contain certain classes of *relevant* information, and with actually identifying specific facts that might answer particular questions or clarify a particular issue.

Finally, the form of the procedures people dealt with were based, at one extreme, on cognitive simulations that modeled fundamental human capabilities and, at another, on engineering solutions concerned exclusively with effective performance. It is clear that not all of the possible combinations of these *variables* could have been represented in the group, but, as will be clear from the obstacles toward the goal of arriving at a consensus, enough of the alternatives were present to ensure diversity!

Perspective Papers

The range of subject matter to be spanned by the Workshop coupled with the differences already noted among the participants made it essential to establish in advance some common base from which the issues could be discussed. Consequently, before the meeting we sent out a variety of materials for background reading. Of course, the book *Linguistics and Information Science* constituted a basic reference both for a review of the literature and for a presentation of the issues. Two chapters on automated language processing prepared for the *Annual Review of Information Science and Technology* (Walker, 1973; Damerau, 1976) covered some of the more recent literature.

To sharpen the focus on issues, we asked some of the prospective participants to reflect on the Sparck Jones and Kay book from the vantage point of the disciplines we had considered in the organization of the Workshop. The results were papers on linguistics, information science, library science, quantitative linguistics, computational linguistics, and complex semantic information processing. These perspective papers also were mailed in advance to those who had been invited and are printed, with some editorial revisions, in this volume.*

----

*A short statement on terminology and translation was presented at the Workshop; it also is included in this book.

Because of the key role they played in the Workshop deliberations, it will be useful to summarize them briefly here. They will be considered in the order in which they are presented in subsequent chapters.

Information Science, by F.W. Lancaster. In his perspective paper on information science, Lancaster emphasizes document analysis, description and retrieval. To provide context, he first examines the activities of information transfer, elaborating on the cycle from creators of information through the publication and distribution process to assimilation by the user community (either directly or indirectly through libraries and information centers) and back to the creators again. Assimilation by the user community may be direct or indirect through libraries and information centers. It is in the information retrieval systems embodied in these mediating agencies that Lancaster explores the relation between linguistics and information science.

"Information retrieval systems are concerned with the acquisition and storage of materials, their organization and control, and their dissemination/presentation to particular user communities." However, the area of organization and control is most relevant for our concerns here, since it establishes the bases for storing documents and for allowing them to be recovered. Lancaster focuses on subject access in his analysis, stressing the parallels between the conceptual analysis and translation phases both of the indexing process and of the search request. The performance of an information retrieval system depends on the quality of the request, the accuracy of its translation into a search strategy, and the effectiveness of the matching process which is limited by the adequacy, accuracy, and exhaustiveness of the initial indexing.

Within information retrieval, Lancaster singles out three major areas in which the application of linguistic techniques, used computationally, can be beneficial. In indexing, they can be considered in relation to extracting words from text, selecting terms from a controlled vocabulary, and using a part of the text itself. In vocabulary control, they can be used to insure consistent representation of subject matter and to bring together terms that are semantically related. In searching, they can contribute to the approximation in an automatic system of the formal structured strategies provided by conventional systems; in addition, they can be use in searching natural language data bases.

In his last section, Lancaster projects the increase in machine readable data bases as leading within 25 years to paperless scientific and technical communication. Online terminals will be used in creating, transmitting, and disseminating documents, in search and retrieval operations, and, broadly, in interpersonal communication. Based on these capabilities, linguistic techniques will assume a critical importance for information systems of the future. Lancaster concludes that linguists and information scientists must collaborate to solve the design and implementation problems these systems entail.

Library Science, by Derek Austin. Austin's perspective from the vantage of library science emphasizes terminological considerations as they affect a large, multi-media, pan-disciplinary library. He begins more generally by examining the factors that bear on the selection of an indexing system. Libraries vary significantly in size from small personal collections to national archives. The types of media they contain may include printed texts, maps, prints, photographs, and audio-visual materials of an increasing variety, produced by both conventional and nonconventional means. The contents included range from a single, highly specific subject field to comprehensive all-inclusive coverage.

The large, multi-media, pan-disciplinary library faces special problems in trying to provide a unified data base accessible by uniform procedures and available to the public in an interactive mode. Austin notes that the number of acquisitions each year is extremely large and that few come with abstracts that might allow current or prospective linguistic procedures to be used. In addition, the non-textual materials are in principle not accessible to such processing. If there is to be a common system for textual and nontextual materials, it will be necessary to have human indexers, Austin argues. However, it is not sufficient to work with term cooccurrence without establishing syntactic relations among them to ensure adequate discrimination.

In response to these problems, Austin has created a system called PRECIS, which he describes in his paper. He has built on certain aspects of linguistic structure in that "(1) the order of terms in input strings, and in the entries generated by a range of transformational algorithms out of these strings, is based by intent upon a subset of the declarative word strings occurring in natural language; (2) the system also employs a number of [natural language] devices, such as machine-produced prepositional phrases, to resolve latent ambiguities in the entries." In the input string, the indexer introduces the component terms of an entry together with codes that specify relations among them. The system produces a full range of entries covering all the headings under which the item should be indexed, including See and See Also references.

Although PRECIS makes use of linguistic structures only in a limited way, it certainly would be possible to increase its sophistication by the use of additional techniques. PRECIS has been used in experiments with materials in languages other than English with definite success. It also looks promising for use as a *translingual switching system* to provide for automatic conversion of input strings in one language into entries in another language.

Quantitative Linguistics, by Wolf Moskovich. Moskovich's perspective paper contains a detailed examination of quantitative linguistics in relation to information science. Quantitative linguistics provides a description of a linguistic system based on estimates of the relative frequencies of the particular phenomena under investigation. Noting that there has been some controversy about where it fits within linguistics, he identifies it as the part of mathematical linguistics that attempts to determine the laws underlying the statistical organization of texts and to reveal structural features of language by analyzing the behavior of the linguistic units in texts. Although the same procedures may be used in information science for document analysis, storage, and retrieval, the primary goal there is to build workable systems. Moskovich explicates the similarities and differences in the use of statistical techniques as applied to text analysis in each area in order to establish the contributions quantitative linguistics already has made to information science and to determine what its future contributions might be.

The major point of intersection of quantitative linguistics and information science is in the description of language phenomena, as opposed to the use of quantitative arguments to resolve qualitative issues or to explain linguistic phenomena. Relevant models are word frequency distributions, measures of sentence length, syntactic complexity, semantic uniformity, and semantic distance; thus, frequency dictionaries and concordances are items relevant for both areas. Moskovich notes that distributive-statistical techniques were applied independently to discover semantic fields in language and to identify similarities among texts for retrieval. Subsequent studies of the associative links among words in different text intervals have had implications both for linguistics and information retrieval, Contributions to information science by quantitative linguistics can be expected to result from its increased attention to characterizing subsets of natural language and to the statistics of word combinations rather than isolated words. Specifically, statistical analysis of the subsets can provide data relevant to scientific and technological *sublanguages*; techniques for studying word combinations can be used to deal with associative term structures. More generally, quantitative linguistics can help information science in three domains: the creation of a lexicographic basis for systems; automatic indexing, abstracting, and document comparison; and the quantitative laws of text organization. Moskovich singles out research in distributive-statistical techniques for major emphasis, pointing out the utility of associative nets of words for information retrieval, the use of machine-constructed thesauruses in interactive searching, and the possibility of complex algorithms for automatic text analysis (morphological, syntactic, and semantic). However, he points out that the ultimate value of quantitative linguistics rests on the significance of the linguistic units that are processed.

Computational Linguistics, by Naomi Sager. Sager begins her perspective paper on computational linguistics by noting that its application to information processing and retrieval has been limited by the difficulty of the problem, the absence of appropriate support software, the lack of detailed grammatical descriptions, and too frequent attempts to develop short-cut solutions. In the area of parsing in particular, that is, in determining the syntactic structure of a sentence, early work resulted in multiple and often spurious analyses due to difficulties in identifying the appropriate attachments for prepositional phrases and other modifiers and the complex problems associated with conjunction, comparison, and ellipsis, among other factors. Grammatical rules do not contain selectional restrictions that constrain which combinations of words may actually occur in a syntactically well-formed construction. Since these restrictions are specific to a particular subject area, to develop procedures that can be used on arbitrary text, it is necessary to be able to distinguish the appropriate constituents but defer building the multiple structures they can form. Another mechanism is the introduction of transformations that operate on the surface parses to produce an underlying representation of the meaning or *deep structure* of the sentence. In this way, it is possible to reduce the number of alternative grammatical forms containing the same information.

Sager describes how these developments are being used in her own work to provide more effective computational linguistic analyses. She believes that the choice of an underlying representation for semantic content at this stage in computational linguistics should be determined primarily by the applicational context. Her transformational decomposition provides a hierarchy of function-argument predications. She has used distributional linguistic techniques to derive patterns of word-class cooccurrence from transformationally analyzed texts that reveal the subject matter for a scientific subfield.

It is possible to apply computational linguistics in a number of areas. Sager's studies have shown that word lists for thesauruses can be generated from function-argument predications to provide semantically sharp and informationally relevant subclasses. Similarly, structured index terms for the sentences of a text can be derived from triples consisting of an operator and its noun arguments that were obtained from transformational decomposition. Further developments should allow deriving more comprehensive patterns that apply to larger amounts of text. In restricted natural language data bases, it also seems possible to format text elements into structures suitable for question answering and statistical analysis. For interactive retrieval systems, computational linguistic techniques would enable the user to direct the search himself but could also be used to increase precision by checking syntactic relations and eliminating false coordinations. In conclusion, Sager notes that although the processing times for sentences still are large, most research groups are using laboratory rather than production models. When the computational linguistic capabilities for mass processing of texts are developed, the computer technology should be available to support them.

Linguistics, by Petr Sgall. In his perspective paper on linguistics, Sgall considers the relevance of linguistic techniques for fact retrieval rather than for automatic indexing or document retrieval. He begins with a detailed analysis of Winograd's language understanding system, identifying the new factors introduced there as explicitly involving the relations between linguistic competence and performance, pragmatics, and reference. Sgall believes that Winograd's work has contributed to a new understanding of the structure of language and the tasks of linguistics in two major ways. First, it provides the basis for a rigorous test of linguistic theories. Second, the *imperative form* for representing knowledge and semantics serves as a model for incorporating those concepts into linguistics.

Previous work in linguistics concentrated on features characteristic of specific languages rather than on those common to them. New developments in linguistic theory, particularly in the areas of performance, pragmatics, and reference, are reflecting a recognition of the importance of language use, with semantic, psychological, and sociological, implications. Studies of the structure of texts are providing a basis for classifying them and for describing their coherence in a way useful for the analysis of dialogs, questions and answers, and other discourse elements. In addition, linguists are beginning to acknowledge the relevance of computing.

However, the major development in linguistics for potential application to information science Sgall finds in the area of linguistic semantics. After considering the work of the transformationalists and the Montague approach in this area, he argues for the direction being followed by the stratificationalists and functionalists. Their work with diagnostic contexts has provided testable criteria and made it possible to establish the units required on different levels of language structure. They find it desirable to distinguish between the linguistic sense or intensional structure of sentences and the logical or cognitive structures that specify their truth value. This distinction would allow handling strict synonymy in the narrow linguistic sense and yet make it possible to identify as equivalent sentences whose cognitive content for a given state of affairs is the same.

Sgall also stresses the significance of topic, focus, and communicative dynamism as they relate to *given* and *new* knowledge in the conversational context. The structure of sentences must be represented so that they reflect the basic conditions of communication. As a result, an adequate characterization of language should include form, function, and the shared knowledge required for understanding. He correlates these notions to the organization and structure of human memory.

Sgall concludes that there are significant developments in linguistics that can be used in fact retrieval, although he acknowledges the problem of extrapolating from the restricted domain of a given experiment to a universe of realistic situations. Constraints can be applied on the users of a system to accommodate to its level of linguistic sophistication, and preediting of texts may be required initially, but the long range results should be natural language programming.

Complex Semantic Information Processing, by Teun A. van Dijk. Van Dijk's perspective paper on complex semantic information processing considers the structure and processing of discourse as reflected in recent work on *text grammars* in linguistics; on the analysis of narratives, conversations, and the like in poetics and anthropology; on the function of language in communication and social interaction in sociology, socio-linguistics, and pragmatics; and on the representation of knowledge structures in cognitive psychology and artificial intelligence. He relates procedures for the assessment of the semantic content of texts to the structure and formation of text abstracts.

In his analysis, van Dijk distinguishes *discourse*, as an empirical, cognitive, and social verbal unit that is physically manifested in verbal utterances, from *text*, which is an abstract theoretical construct that makes explicit the structure of discourse. Documents contain discourses, but understanding takes place only in relation to the text structure assigned to a discourse. The semantics of both meaning and reference are characterized at two levels: micro-structures relate to the structure of constituent propositions and their linear sequences; macro-structures provide a perspective on the structure as a whole. Macro-rules define the mapping between these levels, establishing coherence and connectedness relations; van Dijk considers rules for deletion, generalization, selection, and construction or integration in detail. A macro-structure constitutes a summary of a document, when translated into some conventionally interpretable language, since it defines what is semantically important for a discourse. These summaries themselves may be organized in higher-order macro-structures that define subject domains.

To illustrate his approach, van Dijk provides a detailed analysis of a particular document describing an experiment in social psychology. He presents the micro-propositions underlying the article, the macro-rules that apply to them, the resulting sequence of macro-propositions, two alternative summaries that can be derived from the macro-propositions, and a superstructure for the paper as a whole that places it in the context of similar reports of psychological experiments.

A useful information system along the lines described by van Dijk would require a full morpho-syntax, a meaning and reference semantics, a system of conventional knowledge about the actual world and about relevant possible worlds, and various inference rules to define derivational relations among propositions. Since none of these components is ready yet, any current system can only be partial and theoretical, handling fragmentary parts of discourse, world knowledge, and concepts.

Terminology and Translation, by J. Goetschalckx. Goetschalckx' perspective paper addresses the requirements faced by multilingual documentation systems like the one being established for the European Communities in Luxembourg. Distinguishing two types of such systems, those based on keywords or descriptors and those accessing titles or abstracts, he stresses the need for a terminological control that balances coherence with usage. For the latter system, phrases are essential for retrieval accuracy. However, terms in a phrases are not always in standard form, particularly for inflecting languages, so truncation is required. Procedures are being developed both for automatic truncation and for morphological reduction. The translation requirement complicates the development of an effective system, and when the user needs more than descriptor equivalence, machine translation may be required.

The Challenge Paper

The perspective papers addressed the issues of linguistics and information science from the a variety of disciplines. To complement their breadth and to help focus the discussions at the Workshop, one other paper was solicited, a *challenge paper.* Hans Karlgren provided it under the title "Homeosemy--On the Linguistics of Information Retrieval." He organized his material as a series of disputable propositions, beginning with the premise that

"Linguistics is necessary for the design of future computer-based information retrieval systems." Noting that linguistics is not restricted to natural language processing, he stresses that selectivity of search is the key problem for mechanical retrieval systems. To provide this capability, Karlgren argues that research and experimentation should focus on the application of complex procedures to small files -- to the exclusion of other activities. In particular, he argues against teaching users to accommodate to their systems, but rather to be guided by their difficulties in designing more effective procedures. Retrieving relevant passages from documents can be as difficult for small files as for large.

In charactarizing the retrieval problem, Karlgren distinguishes among three kinds of question answering systems:

order i:	systems with a finite set of questions
order ii:	systems with a finite set of answers
order iii:	systems with an infinite set of answers

These systems differ in the explicitness with which questions and answers and their relations can be specified in advance. Order i systems, which might provide inventory control or travel planning, could in principle be precomputed in the sense that both questions and answers are limited and their relations fixed. In order ii systems, like those for document retrieval, any answer is a subset of the items stored. In order iii systems, like those for fact retrieval, answers can be derived from stored information on the basis of an analysis of the query. Karlgren restricted his consideration to order ii systems as most appropriate for the purposes of the Workshop. In this context, he argued that the key problem is what constitutes a good match between question/requests by the user and answer/offers from the system. Resolving this problem will require looking more closely at the contents of descriptions in their relation to the objects to be retrieved on the one hand and to the requests on the other.

Karlgren distinguishes non-linguistic and linguistic approaches to the solution of the matching problem. The non-linguistic approach requires translating request and offer into a common language and matching according to their identity, with Boolean logic perhaps introduced for more complex computations. While linguistic components may be used in such systems, the matching process itself is not viewed as linguistic. Rather, that process entails reducing variation between request and offer through successive approximations, inevitably with an overall loss of precision. In contrast, a linguistic approach accepts as a premise that exact representation is impossible, but continues by affirming that exact representation is unnecessary and insufficient. Karlgren introduces and elaborates on the concept of homeosemy or similarity of meaning and argues for a metric defining similarities of expressions and specifying the distance or association between them.

Karlgren concludes by stating that linguistics should dissuade documentalists from trying to create ideal retrieval languages, to attack retrieval through language reduction, and to stabilize retrieval languages rather than make them more flexible. Linguistics can provide parsers and other tools, qualitative and quantitative procedures for synonymy and association, and methods for grammatical filtering. In response, linguists should be enjoined by documentalists to study more carefully inexact expressions, shifts of meaning, question-answering, and semantic topology. The results in both directions could be significant and productive.

Donald E. Walker

The Conduct of the Workshop

In the initial planning discussions, the following agenda for the three-day Workshop was proposed: On the first day, the group would establish "where we are", the current state of affairs in linguistics and information science. On the second day, the group would consider "where we should go", identifying a target state of affairs that would constitute a desirable goal. The third day, then, would be devoted to "how we can get there", the research necessary to bridge the gap. However, it was not possible to constrain the participants to this agenda. Our deliberations can be differentiated into three periods that do correspond roughly to the original intent, but the times consumed were different, and our accomplishments did not provide as neat a resolution as we had envisioned. The following summary, although it does not do justice to the richness of the interactions or the subtleties of the disagreements, does convey the substance of what transpired.

The first period did establish a characterization of the state of the art in the relevant areas that does constitute a set of recognizable trends useful in clarifying "where we are". In library science, collections are getting larger and more varied. In terminology and translation, science and technology are becoming internationalized, with demands for shared accessibility of files in a variety of languages from different countries. The automation of scientific and technical information in information science is increasing, with more documents in digital form and a variety of on-line interactive retrieval systems available for accessing them. Computational linguistics is maturing in the capabilities of systems for syntactic analysis (parsing); applications to text analysis are now beginning to be made; and experimental language understanding systems have been built that contain small but *interesting* semantic microworlds. Linguists are now addressing with enthusiasm the problems of *language in use*; performance, as well as competence, is a legitimate area of inquiry; as a result, a more comprehensive view of language is emerging. Finally, in computer science and artificial intelligence, new developments are occurring over a range of diverse areas: on-line, interactive *personal* computers; network linkages with teleconferencing; and cognitive modeling. Because these areas are so dynamic themselves, it is not possible to establish a precisely specified base line from which work on linguistics and information science can proceed. The focus of the Workshop is part of a much larger set of influences and cannot be considered apart from them.

The second period of discussions, rather than specifying directly "where we should go", consisted of attempts to identify the variety of "we's" among the participants. The differences among them in background and orientation (described earlier in this chapter in the section on "The Group") were used deliberately to enrich the interactions, but they had more complex effects on the outcome of the Workshop. It is useful to characterize the major types of protagonists and the goals they reflected here, even at the risk of some caricature, to clarify these effects. The *pan-disciplinary multi-media librarian* concerned with developing an effective system to allow discrimination among masses of diverse materials needed richer linguistic techniques for analyzing noun phrases. The *multi-lingual document specialist* coping with specialized archives in various languages needed spelling rules, transliteration schemes, morphological analysis procedures, and techniques for constructing and analyzing phrase patterns. The *information science experimentalist* working with well-specified corpuses of documents and complex evaluation procedures wanted techniques that had a demonstrable effect on system performance. The *quantitative linguist* with sophisticated mathematical models and processes needed a more precise specification of relevant linguistic units. The *computational linguist* exploring newly developed systems in a variety of application areas wanted more than anything else the time and resources that would allow determining their potential. The *artificial intelligence type* caught up in on-line computer and communication technology and beginning to model a variety of human capabilities -- cognitive and linguistic -- asked only for more technology to allow freedom to explore further. The *theoretical linguist* in his new appreciation for the breadth and scope of language and its use required time to work out the wealth of new ideas he saw emerging from his field.

The Workshop participants exhibited one or another of these identities or blended them in subtle ways. Consequently, it was difficult to involve the group as whole in a coherent discussion; the premises kept changing. People were asked to identify the positions with respect to which they made specific observations, criticisms, and recommendations, so the bases for differences would become more explicit and their effects could be used more constructively. However, the size of the group and the extremity of the divergences in points of view limited the utility of parliamentary strategies of this kind.

This special *group dynamic* had an impact on the determination of "how we can get there" in two ways: First, rather than delineate a single coherent program of research for linguistics and information science, we identified a multiplicity of approaches that responded to the different interests. Second, the efforts to clarify the nature and source of our disagreements became a major preoccupation. As a result, it would be more appropriate to characterize our conclusions as addressing, somewhat more weakly, the issue "what will help". What was most impressive was the intensity and heat of our discussions. It is clear that we had brought together the right people to create *sparks*. However, it also is proper to note that we had perhaps too many sparks and not enough tinder!

Some useful observations can be extracted from our Workshop experience. Since the protagonists identified above all have a proper place in further discussions, a framework for shared understanding must be developed. We need to identify for each approach the responsibilities entailed -- practical and theoretical, explicit and implicit -- that must be satisfied, the underlying technological resources, the scientific concepts on which constructive developments can be built, and the sociological structure of the institutions within which people work. In other words, we need a more precise map of the territory we are exploring.

Increases in contact and communication among people involved in various ways with linguistics and information science are essential. More workshops are in order, but the next ones should have fewer participants, be less heterogeneous, and meet for longer periods. We must get to know each other better in situations that allow us 'ime to arrive at a shared understanding first, but that also let us stay together long enough to build toward some mutually relevant cooperative activities. For credibility among the practitioners, we must address some near-term goals that have demonstrable effects. For credibility among the theoreticians, we must establish a set of concepts that are defensible and productive.

During the Workshop, we discussed a number of proposals for research. One coordinated list of proposals presented a series of projects that applied increasingly sophisticated linguistic techniques to document storage and retrieval. Another more comprehensive list attacked problems in virtually every area of concern that had been reflected in previous discussions. However, it was not possible to discuss any of these proposals at sufficient length to be able to characterize them as the consensus of the participants. Nor were any of them specified in enough detail to provide an adequate basis for their evaluation. Consequently, these lists would not be appropriate to reproduce here. Some of the ideas were anticipated in the perspective and challenge papers; readers who are interested in immediate inspiration are advised to read those documents carefully from that standpoint. A more comprehensive evaluation of the proposals and the preparation of specific project outlines is accepted as the responsibility of the FID Committee on Linguistics in

Documentation. We will provide a next step toward the development of a comprehensive plan to guide research activities for more effective use of linguistics in information science. It is clear that the framework for shared communication mentioned above is essential and that continuations of the personal interactions begun in this Workshop will figure significantly in any future accomplishments.

Linguistics and Information Science: After Five Years

Karen Sparck Jones and Martin Kay provided the point of departure for the Workshop in their survey, *Linguistics and Information Science*. Both of them joined in our deliberations as participants, contributing significantly to our discussions. In addition, after the Workshop was over, they got together and wrote a postscript to review the major developments since the completion of the manuscript for their book in 1971. This paper, which focuses specifically on linguistics and information retrieval, provides a fitting concluding chapter for this book.

Within theoretical linguistics, the authors note that semantic problems associated with quantifiers and related logical issues have led some transformational grammarians to look for abstract deep structures. An interest in functionalism, associated with the notion of speech acts, has led to an interest in presuppositions, performatives, and other pragmatic phenomena associated with language use. As a result, the distinction between the notions of competence and performance seems to be breaking down, and linguists are no longer able to limit themselves to consideration of isolated sentences. Within computational linguistics, the computational metaphor has become increasingly compelling in guiding the development of models for the production and understanding of utterances. Although current systems are limited to micro-worlds, they demonstrate significant progress of potential relevance to information retrieval.

In information retrieval, the major development has been the growth of on-line search systems for very large data bases. While the effectiveness of these systems has increased substantially, they reflect, if anything, less use of linguistic capabilities. Rather, computational power provides flexibility by allowing complicated specifications to be made. In addition, statistical techniques and automatic weighting schemes are being used with increasing success. However, there are still too few relevant experiments to enable any of these procedures to be evaluated.

Sparck Jones and Kay are not discouraged by their conclusion that the connection between linguistics and information retrieval is not greater now than it was when their book was written. They account for this fact in relation to the sharp distinction in scale between the concerns of the two disciplines. However, they note that linguists are becoming interested in larger units of discourse and they believe that retrieval systems soon will have to retrieve information units smaller than documents. Consequently, they expect greater collaboration in the future, a conclusion clearly echoed by the Workshop deliberations -- although the steps toward that goal still remain to be defined in research terms.