Information Retrieval and the Philosophy of Language

D. C. BLAIR
The University of Michigan, Ann Arbor, Michigan, 48109-1234

This discussion takes the position that information retrieval systems are fundamentally linguistic in nature—in essence, the languages of document representation and searching are dialects of natural language. Because of this, the discipline of the Philosophy of Language should have some bearing on the problems of document representation and search query formulation. The philosophies of Austin, Searle, Grice and Wittgenstein are briefly examined and their relevance to Information Retrieval theory is discussed.

Received December 1991

Information Retrieval systems are fundamentally linguistic: the content or context of documents must be described, and the inquirers' needs for documents must be expressed. These descriptions and expressions are most frequently articulated in free or controlled vocabularies that have some of the same characteristics as natural language. As a consequence, the processes of document description or request formulation must be strongly related to the processes of description and inquiry in natural language. This is not an original observation, the fields of linguistics and formal logic have had an early and continuing impact on Information Retrieval theory [see, e.g. Cooper, Sparck Jones, and more recently, van Rijsbergen].10,25,28 But recently Information Retrieval theory has begun to be influenced by past and present theory in the Philosophy of Language.6,13 The rationale for this shift is clear: the more we understand about how language works, the better we can understand how to describe and ask for documents. The Philosophy of Language, though a difficult, and at times obscure, discipline, has a lot to say about how language works. It is the purpose of this paper to show how, first, Philosophy of Language has already helped Information Retrieval research, and, second, how it may continue to do so. To this end, this paper will discuss three specific Philosophies of Language and how they relate to Information Retrieval research: the theory of Illocutionary Acts of Austin and Searle; Grice's theory of conversation; and Wittgenstein's Ordinary Language Philosophy.

Information Retrieval and Illocutionary Acts

Traditionally, the Philosophy of Language has been primarily concerned with the propositional content of language—the nature of factual discourse. Most battles in the Philosophy of Language have been waged over the grounds of truth, that is, how it is that we can assert that certain things are, in fact, the case. Although the evaluation of the propositional content of language has had an influence on the design of computerized information systems—primarily 'fact retrieval' or question-answering systems—it has not had a significant influence on information retrieval design [an exception to this is Cooper].10 But language is not solely concerned with factual assertions. John Austin showed that there was a class of linguistic acts that are not governed primarily by truth conditions.2 In a large number of cases language could be used to do things: that is, instead of merely describing what had occurred, language could also be used to make things happen; Austin called this the performative nature of language. I can say:

(1) a. I'll pay you right back.
   b. I name this ship the 'Norton Sound'.
   c. Finish the report before tomorrow's meeting.
   d. Bill's a better worker than Bob

When we say things like this (make a promise, christen a ship, give an order, make an evaluation) we aren't so much talking about something, we are in fact doing something with our statement. If a reasonable individual in a normal situation promises to meet you tomorrow for lunch, then by virtue of that statement, he has made a promise. It would make no sense to ask, 'Did he really make a promise?' Of course, he could break his promise, but that would not change the fact that a promise had been made. What makes the promise, or any other performative, 'work' is a set of 'felicity' conditions—a set of normal circumstances—that are presupposed by the performative act. For example, the individual must be the kind of person who can promise (e.g. an infant can't promise), the circumstances must be appropriate for the promise (e.g. I can't promise you something if I am all alone and no one hears me), I can't promise what is beyond my power (e.g. that I can make you the Duke of Kent), etc. Performatives work within the broad context of personal and social conventions, and often require an institutional context in which to be successful. For example, only a minister can declare a man and woman 'man and wife'; and this can only take place in an appropriate ceremony, with a prescribed number of witnesses, following local and federal statutes that specify who is legally eligible to marry. Further, the man and woman must be willing to marry each other, etc.

Some illocutionary acts are very formal and quite institutionally dependent (e.g. legal contracts, or orders within 'chain of command' in the military) while others are less formal (e.g. ordering a drink at a bar, or promising to have lunch with a friend). Nevertheless, the most important characteristic of illocutionary acts is that they, by virtue of being uttered in the appropriate circumstances, cause something to happen (a promise to be made, an order to be given, etc.).
A taxonomy of Illocutionary Acts

Although both Austin and Searle proposed that Illocutionary Acts fell into several specific categories, Searle’s taxonomy, while similar to Austin’s both in number and type, has had the greater impact on recent linguistic theory and information system design (Searle calls these acts ‘Speech Acts’, but the descriptions ‘Speech Acts’ and ‘Illocutionary Acts’ can be used interchangeably). The principal difference between Austin’s and Searle’s classification of Illocutionary Acts is, according to Searle,23 that Austin’s classification is really a classification of illocutionary verbs, while Searle’s is a classification of illocutionary acts. Further, while Austin was uncomfortable grouping his illocutionary verbs into categories, Searle has been quite insistent that his classification is definitive and final. In fact, Searle limited the number of different illocutionary acts in direct response to Wittgenstein’s insistence that there are ‘countless’ kinds of sentences, or what he called Language Games. While it is not clear that Illocutionary Acts and Language Games are the same kind of phenomenon, Searle felt that they were similar enough to take Wittgenstein’s claim about Language Games as being, ceteris paribus, applicable to Illocutionary Acts, too.

Briefly, according to Searle, there are the following kinds of Illocutionary Acts:

- **Assertives**: in which we tell others (truly or falsely) how things are. E.g. ‘Bill was at Mary’s party last Friday.’
- **Directives**: in which we attempt to get others to do things. E.g. ‘All hands on deck!’
- **Commissives**: in which we commit ourselves to doing specific things. E.g. ‘I’ll send the report to you tomorrow.’
- **Declarations**: in which we bring about changes in our world by our utterance—in short, saying makes it so. E.g. ‘I now pronounce you husband and wife.’
- **Expressives**: in which we express our personal feelings and attitudes. E.g. ‘You did a great job!’

Searle noticed that there is some overlap between Assertives and Declarations, and proposed a sixth category, or sub-category, called ‘assertive declarations’. These are Declarations that are based on the assertion that something is or is not the case. For example, when the referee in a basketball game cries ‘Foul!’, he declares a penalty and asserts that some specific event has occurred. The difference between this category and ‘pure’ declarations is that for the Assertive Declaration the speaker can lie, but for the ‘pure’ declaration the speaker cannot lie. The principal difference between Austin’s and Searle’s classification is that Austin has no category that corresponds to Searle’s ‘Assertives’. There are rough parallels between the other categories, though.

Illocutionary Acts and Information Retrieval

Blair9 has proposed a document indexing structure based on Austin’s taxonomy of Illocutionary acts, but the theory of Illocutionary Acts has had its greatest influence in electronic messaging systems. Such systems, since they usually involve the transmission of textual information, can be thought of as a special case of information retrieval. The best-known application of Illocutionary Acts to electronic messaging is the COORDINATOR system.12,20,31 The structure of the COORDINATOR is based on the observation that messages are not always independent, but are often parts of ‘conversations’ that, together, carry out some activity. If the ‘activities’ of language can be placed under five well-defined Illocutionary Acts, as Searle believes, then, by inference, messages or documents can be similarly classified (of course, a message or document may comprise more than one Illocutionary Act, which means that such classification may not always be straightforward). What the COORDINATOR does is to structure electronic messages according to the Illocutionary Act under which it falls (it also urges the sender to limit his/her messages to those that perform a single Illocutionary Act). For example, if an individual makes a request, he identifies his message as a ‘request’ and the system then imposes a general ‘request structure’ on that and all subsequent messages that are part of this transaction. That is, individual A may make a request to B, who, in turn, may promise to fulfill A’s request, or make a counter-offer of how he, B, would prefer to fulfill A’s request. A, then, could accept B’s offer or make a counter-offer of his own, putting the onus on B to continue the dialogue. Of course, at either time one or both of the individuals can cancel the dialogue, or conclude it. The COORDINATOR contains procedures to manage this interchange efficiently. It will prompt the initiator of a request for a ‘respond by’ date and/or a ‘completion’ date, and then prompt the receiver of the request to reply or finish by that date. It also organizes and presents the ongoing ‘conversations’ (requests, promises, offers, what if, or questions) in a structured format so that the individual using the systems can see exactly how each of his conversational activities stands in terms of completion or need for further dialogue. An individual could be presented with a screen on his computer/workstation that summarizes both the responses due to him as well as the responses he is required to make. It can also present a complete history of all the messages exchanged in a given ‘conversation’.

A more ambitious application of Speech Act Theory to electronic messaging is the work of Kimbrough et al.16 Kimbrough has been working on a Formal Language for Business Communication originally based on the illocutionary logic of Searle and Vanderveken.24 Kimbrough has since abandoned illocutionary logic and has developed a method for capturing the propositional attitudes of Speech Act Theory in first-order predicate logic.17 This modelling language has been applied primarily to electronic messaging systems.

While electronic messages are often thought of as ‘ephemeral’ documents and are not routinely kept for any length of time, electronic messaging systems increasingly are being used to undertake transactions which may be important to keep. Consequently, it will be necessary to store the messages/documents created by them in a more traditional information retrieval system. In that case, it would be useful if the document representations were to be based on the Illocutionary Acts that were used to structure them when they were messages. That is, the taxonomy of Illocutionary Acts becomes a classification scheme for the representation of performative messages, and the links that were made
with other messages that were used to perform the same act become useful ways of clustering the messages when they are stored as documents. The description of the subject content of these documents, as well as the recording of important contextual information (such as the author's name, date the message was written, etc.) may still remain important parts of the document representation, but the taxonomy of Illocutionary Acts and the linking of documents according to their mutual participation in a particular activity are important additions to our theories of how to represent documents for retrieval.

Traditionally, Information Retrieval theory has been applied primarily to scholarly articles and related journal communication; but even this kind of organization can be seen often as implicitly following certain types of Illocutionary Acts. For example, the citations at the end of a scholarly article are often used as links to previous articles that are, in the author's opinion, relevant to his/her work. The inclusion of a citation in the bibliography of the article is, from an illocutionary viewpoint, an ‘assertive declaration’; that is, the author declares that the citation(s) are part of the bibliography of the article, and asserts that the articles they refer to are relevant to the citing article. It is also clear that scholarly articles routinely contain expressives (e.g. evaluations of previous research), directives (e.g. ‘see’ or ‘see-also’ references), and even commissives (e.g. ‘This work will be continued in a later paper’).

Future applications of Illocutionary Acts

There are several ways in which the basic notion of Illocutionary Acts can be expanded to assist in information retrieval. In the first place, the application of Illocutionary Acts as an organizing principle for information retrieval should not be limited to Searle's taxonomy. Searle's rigid classification of Illocutionary Acts, it must be remembered, is primarily useful for maintaining distinctions important to the Philosophy of Language. Such distinctions may not be as crucial for information retrieval. In fact, casual inspection quickly shows that, for example, all promises are not the same. A promise between two business associates to have lunch is clearly not the same as a promise (that is, a contract) by a company to provide specific products or services to a client. The latter would be, at the very least, a binding legal obligation, while the former would not be. As a result, if messages were used to carry out each of the above activities, they would certainly require different kinds of information to represent those documents. For example, to execute a business contract, specific documents must be drawn up, and certain empowered individuals must authorize or witness the execution of them. Further, the messages that document a legal contract may be used quite differently than a message promising to have lunch with someone. The documents used to execute the contract would probably be used to demonstrate that the contract had been executed in good faith and according to proper, binding procedures. The message promising to have lunch might only be used to establish that two individuals probably met at a particular point in time, but could not indicate much beyond that. Both could be important from a legal point of view, and both are promises, in the strict sense; but the structure of the acts which they perform differ markedly. This difference would need to be taken into consideration in the representational scheme of the information retrieval system that manages them. Similar variety could be found in the other categories of Illocutionary Acts (Assertives, Directives, Declarations and Expressives), and even among philosophers there is no consensus that Searle's taxonomy is final (Bach and Harnish give a detailed description of this variety, breaking each category down into more specific types). The originator of Illocutionary Act theory, John Austin, believed that even his own classification of such acts was very rough and not rigorous, and Wittgenstein believed that he might be useful to use a performative taxonomy to represent documents for retrieval (although there certainly remains a great deal to be done working out the details of an indexing or representation scheme based on Illocutionary Acts). It also points to the importance of information external to the text of a document for representing or describing that document for retrieval. Traditionally, document representation has been thought of primarily as representing the intellectual content of documents — what they are about. The theory of Illocutionary Acts, when applied to information retrieval, shows that documents not only have a content, but they have a use — in fact, they may have multiple uses (for example, a document could be used to execute a contract, but could later be used as evidence to support the assertion that the contract was valid); and these uses of a document may not be entirely deductible from the text of that document. The reason for this is that the successful performance of an illocutionary act is predicted on the satisfaction of certain more or less specific felicity conditions, and usually the satisfaction of the felicity conditions is part of the milieu of the document, not part of its text. For example, the drawing up and signing of a contract (promise) does not necessarily execute the contract for the reason that only certain individuals are empowered to execute the contract. In this case, an individual can be said, in effect, to promise something, whereas it could be shown in a court of law that although he made the ‘right noises’ he did not actually make a promise (since he was not empowered to do so). Since the satisfaction of felicity conditions is an important precondition for the execution of an Illocutionary Act, then it stands to reason that the felicity conditions themselves may also play an important role in any document representation scheme based on Illocutionary Acts (for example, if only specific, empowered individuals are allowed to execute contracts, then an information retrieval system that stores contracts should have some way of maintaining an authority list of empowered individuals and be able to match it to the names of individuals who have executed stored contracts).

We mentioned before that even scholarly articles can be seen as conducting Illocutionary Acts, and, clearly, there are ways to extend such analysis. First of all, the
declarations, commissives, expressives, etc. found in scholarly articles can be broken down further. For example, we stated that citations were implicit assertions that the articles or books cited were relevant to the citing article. But certainly there is more than one way in which a previous article or book can be relevant to the citing article, so the implicit assertions of the bibliographies are not all asserting the same thing (Hodges describes 14 different kinds of relation that a citation could imply).

Wilson has demonstrated that there may be a way of classifying scholarly writings that has some interesting similarities to Illocutionary Acts. He has shown that the traditional 'form subdivision' of the Library of Congress is similar to what has been known in literary studies as genre, that is, kinds of literature. Wilson makes the point that the notion of genre has applications well beyond just literary studies, and has characterized genres as 'long Speech Acts' [personal communication]. 'Any field of human activity is likely to develop a repertory of 'definite and relatively stable typical forms of construction' characterizing linguistic communication in the field; 'A particular function (scientific, technical, commentarial, business, everyday) and the particular conditions of speech communication specific for each sphere give rise to particular genres, that is, certain relatively stable thematic, compositional, and stylistic types of utterances. 'Nonliterary genres have not been the subject of much serious study, but it becomes clear on reflection that more or less well-settled conventional types are to be found throughout the world of text production, not merely in the literary section of that world'[p. 37. Wilson quotes Bakhtin]. Clearly, then, there are issues in the control of scholarly texts that are similar to, and may be informed by, the notion of Illocutionary Acts.

**Brute facts vs. institutional facts**

A final distinction that Searle makes (originally proposed by Anscombe) is that between 'brute facts' and 'institutional facts' [see Blair]. Brute facts are similar to the largely physical sense data which we call 'facts' in the natural sciences. The reductionist nature of scientific investigation assumes that all scientific study is a study of brute facts, and as much as the study of language aspires to scientific status it too, has been dominated by the study of the brute facts of language – syntax, word frequencies, statistical models of word/phrase occurrence, etc. Yet, in spite of the superficial rigor that the tabulation of such linguistic brute facts attains, it can never do what it purports to do, it can never be the basis for any reasonably complete semantic analysis. The reason for this, according to Searle, is that meaning in language is based on institutional facts, not brute facts. Meaning in language is best accessible through an analysis of conventions and human activities that are inextricably tied up with human institutions. But although these institutions may have some brute or statistical facts about them, Searle is quite insistent that institutional facts are not derivable from brute facts in any complete way. In this, Searle would find agreement with his predecessors, Austin and Wittgenstein; and it was Wittgenstein who put it so succinctly, 'Only in the stream of thought and life do words have meaning' [Wittgenstein, para. 173]. Yet, if this is the case, then we have a real problem in information retrieval, for clearly the words and phrases that we use to describe and search for documents have been, for the most part, taken out of 'the stream of thought and life'. In fact, what is most readily accessible for analysis in computerized Information Retrieval systems are the 'brute facts' of documents: the frequencies by which index terms are assigned to documents; the frequencies in which terms co-occur in their assignment to documents; and for full-text retrieval systems: which words occur in the text of documents; which words co-occur with others in the same document, the same sentence, in the same paragraph, and within a specified proximity of other words. On this simple foundation of brute facts we have built remarkably complex models of document representation. But have we captured the 'meaning' of the documents represented by these simple facts? If Searle is correct, we have not – nor can we. No number or complex combination of brute facts can be produced to give us the 'meaning' of a document – where the 'meaning' of a document would include such things as its subject, intellectual content, context, use, purpose or links to other documents. The 'meaning' of a document is underdetermined by the brute facts of that document. Of course, this is not to say that word or index term occurrences have no relation to the meaning of a document. There is a relation, but it is an adumbrative one, where the occurrences of words in a text (or some complex metric based on them) only hint at the 'meaning' of the document.

The obvious question, though, is: 'Do we really need the "complete meaning" of a document in an information retrieval system?' Perhaps not. The only way we could tell whether 'semantically sparse' document representations were good enough is to conduct frequent, extensive tests of retrieval effectiveness on large, operational systems. This, unfortunately, is not done very often – primarily because such tests are so costly. But this may also be a reason for applying the theory of Illocutionary Acts (as well as other theories of language) to document representation. That is, if Illocutionary Acts are good, though perhaps incomplete, models of language usage, then using them as a framework for document representation guarantees, at the very least, that those document representations will function more like natural language than document representations that are not tied to any theory of language use. It is a reasonable assumption that document representations which function more like natural language will work better than those that do not, even if we do not have the empirical evidence based on tests of retrieval effectiveness to support it. Theories of natural language usage, then, may become the theoretical foundation that we need in information retrieval. This is not to say that there won't be other theories that will help, too. But these theories of natural language use may become the theoretical bridges of information retrieval that, like the theories of astronomy and quantum physics, allow us to make reasonable advances in document retrieval theory in between the rare empirical tests of large-scale system effectiveness.

**The inferential nature of language**

If models of language use have some relevance to information retrieval theory, then it is no great in-
intellectual leap to see why they are applicable. In short, the language of document representation and searching is a 'dialect' of natural language, and the searching and indexing processes are, fundamentally, processes of communication between indexers and searchers. That such communication is awkward or stilted does not change the fact that it is, fundamentally, communication. As communication, the language of information retrieval must conform to the principles of conversation. For an understanding of this area of linguistic theory we need to turn to the work of Paul Grice. Until Grice's work, the ascendant theories of meaning in language were, explicitly or implicitly, theories of coding (see, e.g., Eco). That is, the way that one individual understood another was for the listener to hear the words/phrases of the speaker and then to 'decode' those words/phrases into some kind of semantic content. Different theories of meaning were distinguished primarily by how they asserted that this decoding took place. Grice showed that the decoding theories could not account for, among other things, how a hearer could tell when a speaker mis-spoke: for example, when the speaker made a mistake or lied (that is, simple decoding theories would always take language literally, something that is clearly not the case in natural discourse). Grice showed that this phenomenon could only be accounted for if the hearer inferred some or all of the speaker's meaning and intentions prior to the speaking of the words in question. In short, Grice was able to show how it was possible to convey an unambiguous idea with an ambiguous sentence.

Grice never said that no decoding takes place in natural discourse; understanding is a process of both decoding and inference. The inferences that individuals make about a speaker's meaning are based on the recognition of subtle contextual and behavioural cues (e.g., I know that although Mary says 'Thanks!' the tone of her voice, her manner and what transpired before, has told me that she is being sarcastic). Such inferences are not generally available to the inquirer who uses an information retrieval system or to the indexer who prepares documents for retrieval. Consequently, the language of information retrieval is missing much of the dimension of inferential communication, and reduces to a system of literal semantic transactions; hence, most theories of document representation or indexing are implicitly coding theories. What effect does this have on the process of information retrieval? It's not clear, without, again, many tests of retrieval effectiveness. But, lacking such tests, and accepting the hypothesis that making the language of information retrieval more like natural language is a good thing, it appears that information retrieval languages - to become more natural - should have two components: a set of descriptors and rules to assign meaning to them; and, a system of contextual reference that provides information that would permit the inquirer to infer other, non-literal meanings from the words which describe the documents in the system. Just what this system of contextual reference should comprise is not clear, although some preliminary suggestions were made [see Blair, pp. 183ff.]. It may also be the case that the system of contextual reference may be composed mostly of 'institutional facts'. For example, it may be that for an inquirer to retrieve documents effectively on an information retrieval system used in litigation support, there may need to be included in the indexing structure information specific to that lawsuit, e.g. the titles and affiliations of individuals named in the documents; significant dates, such as when contracts were signed, the time periods in which individuals held significant positions in the involved institutions; the professional and perhaps casual relationships between individuals named in the lawsuit; etc. All of this would be sensitive to the specific context of retrieval. The inherent nature of natural language implies, therefore, that any theory of document representation/query formulation that does not take this inferential dimension into consideration may be less effective than a theory that does.

The cooperative principle

Also of interest to information retrieval is Grice's notion of conversational cooperation. For Grice, our conversations are characteristically cooperative efforts. Because of the inherent ambiguity of meaning, in order for the hearer to make any inferences about the speaker's intended meaning, the hearer must assume that the speaker is honestly trying to communicate in good faith. This leads the hearer, according to Grice, to assume several quite specific things about the speaker's intentions, namely:

Maxims of quantity
(1) The speaker should be as informative as required.
(2) The speaker should not be more informative than required.

Maxims of quality
(1) The speaker should not say what he believes to be false.
(2) The speaker should not say that for which he lacks evidence.

Maxim of relevance
The speaker should be relevant.

Maxims of manner
(1) The speaker should avoid obscurity of expression.
(2) The speaker should avoid ambiguity.
(3) The speaker should be brief.
(4) The speaker should be orderly.

In so far as an information retrieval search is like a conversation - with the inquirer making requests and the retrieval system 'answering' with sets of documents which match the requests - it stands to reason that these maxims of cooperation hold up in the search process. Note an important distinction, though; Grice's principle of cooperation does not assert that these maxims are always upheld in conversation, he asserts the more subtle point that they are assumed to be upheld in conversation. That is, unless there is evidence to the contrary, the listener assumes that the speaker is being as informative as necessary, is not lying, is relevant, brief, orderly, etc. In information retrieval, then, it is likely that the inquirer submitting search queries to a system and receiving sets of documents in return assumes that his search queries are being dealt with cooperatively. But is this the case? It is often hard to tell. Unlike an ordinary conversation where we have continual evidence of cooperation, or at least the means by which to test it, in the information retrieval process it is more difficult to determine whether the maxims of cooperation are being fulfilled. In fact, we
can only easily detect the violation of two of the maxims of cooperation: the second maxim of quantity: the speaker should not be more informative than necessary; and the maxim of relation: the speaker should be relevant. The first maxim is obviously violated whenever the inquirer receives excessively large sets of retrieved documents in reply to his queries, and the second maxim is violated whenever the inquirer receives retrieved sets with large numbers of non-relevant documents along with relevant documents (n.b. the retrieval of no relevant documents is not necessarily a violation of the maxim of relation – for the obvious reason that there may be no relevant documents on the retrieval system). The other seven maxims of cooperation are very difficult even to challenge (much less substantiate) during the retrieval process. Why is this? Just think how we would challenge whether one of these maxims is being upheld during a normal conversation: we give the speaker feedback about his/her conversation. For example, we might say, ‘Aren’t you beating around the bush here?’ if we think he is not being brief, or, ‘That’s pretty ambiguous’, if he appears to violate the second maxim of manner, or, ‘Can’t you just say it in plain English?’ if he is being obscure, or, ‘Are you sure that’s right?’, if he violates the first maxim of quality, or, ‘What makes you believe that?’ if he violates the second maxim of quality, etc. Our language is replete with phrases that would permit us to challenge any of the maxims of cooperation during the course of conversation. It also gives us linguistic ways for answering these challenges. For most information retrieval systems, though, we cannot ask these kinds of questions; and since we cannot ask these kinds of questions, we cannot challenge the cooperation of the system. Because the non-cooperative conversation is the exception rather than the rule in ordinary discourse, and since we do not have the ability to assess the cooperation of the information retrieval process, it is natural that inquirers usually will assume that retrieval systems are more cooperative than they actually may be. This explains why, in spite of the uniformly low reported levels of retrieval performance, inquirers generally do not express much dissatisfaction with large, operational document retrieval systems. In short, inquirers typically assume more cooperation – that is, conscientious retrieval – than is probably the case. This may also explain why the evaluation of large scale, operational retrieval systems is not treated with the urgency that it deserves, and why empirical tests which show low levels of retrieval effectiveness, even when rigorously carried out, are often met with scepticism or even disbelief.

If information retrieval is really a form of conversation, then Grice’s maxims of cooperation become heuristics for effective information retrieval: Maxims of cooperative retrieval. Consequently, the challenge of information retrieval research is to build into retrieval processes effective tools that the inquirer can use to challenge and assess the cooperativeness of the document retrieval process. This will not be an easy task, but it is better to work on a difficult important problem than to work on a more tractable but less significant problem.

Wittgenstein

Wittgenstein’s Philosophy of Language defies simple explanation, and there already exist both explanations (especially by Pitkin23) and a lengthy discussion of its application to information retrieval. So the discussion here will be brief. But even though Wittgenstein’s philosophy of language is difficult to summarize, the following two quotations encapsulate the essence of Wittgenstein’s understanding of language.

For a large class of cases – though not for all – in which we employ the word ‘meaning’ it can be defined thus: the meaning of a word is its use in the language [Wittgenstein24, para. 43].

We don’t start from certain words, but from certain occasions or activities [Wittgenstein24, p. 3].

Where Searle saw simplicity and only a few types of language use, Wittgenstein saw a welter of complexity. For Wittgenstein, language is primarily a set of tools which we use to engage in certain specific activities. We learn how to use language in the same way that we learn how to use tools or implements – not by definitions and explanations, but by having the appropriate usage demonstrated or shown to us by means of what Wittgenstein called ‘perspicuous examples/representations’ (übersichtliche Darstellung); and the only place where we can see language being demonstrated correctly is in the ‘occasions or activities’ in which it is embedded. Language does have a structure, but it is not a grammar in the traditional sense; and this structure is dynamic like a game being played, rather than a set of instructions to be followed. The structures of language use are embodied by what Wittgenstein called ‘Language Games’, and the occasions or activities in which these games are embedded are the intensely human activities that he called ‘Forms of Life’ – what he sometimes referred to as our ‘Natural History’. Our language, then, is inextricably caught up in the things that we do, and we must understand how we participate in these activities before we can understand how language is used in them. In this sense, Wittgenstein would appear to favour Grice’s inferential theory of meaning rather than a more traditional coding theory. In other words, like Grice, Wittgenstein would agree that we often know a lot about what a speaker will say before he says it. We have a complex set of ‘criteria’ that help us disambiguate both language meaning and the speaker’s intention.

But the point of Wittgenstein’s analysis of language was not just to try to show how it worked, but to also show how its misuse could lead to certain systematic misunderstandings – what he called ‘diseases of thinking’. The primary cause for these problems was that those who looked at language (mostly, philosophers) tried to examine it independently of its usage, and independently of the activities in which it was ‘at home’. The scientific method of Wittgenstein’s day was primarily and extractive one, a method which isolated the object of study and removed it from its natural context. While such a method may have success in some scientific pursuits, for Wittgenstein it could only be misleading in the study of language. Further, the reason problems in language can lead to mistakes in thinking was, according to Wittgenstein’s theory, because language was not the product of thought, but the vehicle of thought:

When I think in language, there aren’t ‘meanings’ going through my mind in addition to the verbal expressions: the language is itself the vehicle of thought” [Wittgenstein25, 1953, para. 329].
Language is the means by which the speaker works his thought out—it is the set of tools and the work area which are the foundation of his thinking. The speaker's ability to think clearly is as reliant on his mastery of language as the artist's ability to express himself is reliant on the paints and brushes he has and his ability to use them.

Although we have described Wittgenstein's Philosophy of Language in the briefest possible terms, we can see some broad implications for information retrieval. In the first place, the language of document retrieval, like ordinary language, must have its meaning grounded in activities. Consequently, there won't be one way of describing a document, but a variety of ways, each based on the activity that uses the document in question. Thus, information retrieval systems are activity-specific. They, like their language, are dependent on the activities that they serve. The role of the indexer or the designer of indexing algorithms is to relate the usage of the terms which represent the documents to the usage of those words in the activities that employ those documents. As a result, the study of information retrieval can be thought of as the study of information in context. We cannot separate the design of information retrieval systems from the activities in which they are embedded (some Information Retrieval researchers have begun to explore this connection between document description and activities; see Oddy et al.20). Further, if our language is truly a vehicle of our thought, then the kinds of thinking we can do in the process of searching for documents will be strongly dependent on the language used to represent those documents. The level of complexity or refinement for working out and expressing our information needs will be ineluctably limited to the level of complexity or refinement of the language of representation available to us.

Conclusion

This discussion has been an attempt to outline, however briefly, some parts of the Philosophy of Language that have bearing on the task of Information Retrieval. Clearly, information retrieval systems are in large part linguistic systems, so it has been the implicit thesis of this discussion that theories of language use or meaning are, mutatis mutandis, theories of document representation or query formulation. The natural question at this point, then, is why do we need any theory of information retrieval at all? Why can't we just build information retrieval systems, changing and adjusting them as we go along until we get good systems? The answer is, that because of the linguistic nature of Information Retrieval, there are simply too many degrees of freedom in design for us to arrive at good designs haphazardly. But if the language of Information Retrieval is really based on natural language (here, English) in some non-trivial way, then it stands to reason that any theories of how natural language works will help us to understand the linguistic aspects of Information Retrieval. To build a language for information retrieval without considering how natural languages work is presumptuous at best, impossible at worst. As Thomas Kuhn wrote,18 one of the necessary characteristics of a scientific discipline, as opposed to a non-scientific or pseudo-scientific one, is that its failures are informative, they exclude certain avenues of research and encourage others. The Philosophy of Language can help to make our failures in information retrieval design more informative than they presently are.

But although theories of natural language use can be helpful for understanding of the linguistic aspects of information retrieval, there may be aspects of the language of information retrieval that have no clear correlation with the processes of natural language. Specifically, because of the enormous size of their document collections, some current retrieval systems are creating linguistic universes that are unprecedented. For example, our language was never meant to make the kind of subject distinctions that it is being called on to make in large-scale systems (one merely needs to look in the subject catalogue of a large research library to see how inadequate most subject classifications are for making fine distinctions in large document or book collections—and these research collections are usually indexed by professional cataloguers, a luxury that most commercial information retrieval systems do not have). It may be the case that it is simply impossible to make fine distinctions of intellectual content on large document retrieval systems. If this is the case, then the kinds of systems we can build and the levels of effectiveness that we can expect from them may have to undergo substantial reconsideration (for example, if fine distinctions in the intellectual content of documents must be made, then it may be more profitable to develop strategies to keep the collection small, rather than letting the collection grow and trying to make finer distinctions in the indexing language).21 Whatever the case may be, we probably have not treated the problem of document representation with the care and intensity that it deserves. If the languages of document representation and searching are really dialects of natural language, then we should expect that any theory of document representation and searching will be no less complex than the theories of language meaning and use are.

REFERENCES

9. D. C. Blair and M. E. Maron, An evaluation of retrieval effectiveness for a full-text document retrieval system:
information retrieval and philosophy of language


Book Review

RAJ JAIN

A friend of mine once said that 'common' sense was pretty rare in the computing area. A good dose of common sense is a common factor in nearly all myfavoured computing texts. In order to measure this book's performance we must define our criteria. To a deal of common sense we may wish to add the criteria of comprehensibility, clarity and interest.

The book scores well on common sense: for example, a number of checklists are presented to highlight common problems encountered. The author reinforces the view that simple techniques are always better than complex ones where the simpler ones will suffice.

The book is admirable in its comprehensiveness. A wide range of theoretical techniques are presented, but the large number of practical examples prevents the treatment from becoming dry or sterile. The book is divided into six sections comprising an overview, techniques and tools, probability and statistics, experimental design and analysis, simulation and queuing models. Each section concludes with a section pointing the reader to further sources in that area. The author claims that no other text covers this range of topics in a single treatment, and the reviewer would concur with this view. Certainly the book has revealed a number of techniques new to me.

The text compares favourably for clarity with texts on probability and statistics encountered by the author. A text of this type cannot assume prior knowledge of such a broad range of topics. The text requires a background in computer systems but no great prior knowledge of statistics. The structure is of the 'tell them what you're going to say, say it, and tell them what you said' type, based on the six principal sections. This aids the clarity of what might otherwise become an unstructured collection of methods.

Interest is maintained through the author's writing style together with a large number of examples and case studies. Overall, the book is recommended for its interesting and comprehensive treatment of the subject. However, there are a number of limitations which should be mentioned.

The first arises from the broad scope of the text. Some subjects are treated in a rather superficial way, for example chapter 29 provides a list of commonly used distributions, but with limited explanation. It is not always apparent where to find an appropriate technique if you dip into the book without looking for a specific named technique.

Perhaps the most obvious flaw is the need for four pages of tabulated errata, particularly fifty errors in the Author Index. As an author myself, I understand the difficulty in ensuring that a correct and complete manuscript reaches publication. I also respect the author's integrity in providing such a full and pains-taking errata section. However, the basic integrity of the text should be better, and it is inconvenient to keep checking the information contained in the main text against the errata.

Taken as a whole, these limitations notwithstanding, the text is recommended as a comprehensible and comprehensive treatment of the subject, written in an interesting way. Let's hope that it sells well enough to require another print run, so that those niggling errata can be incorporated into the main text.

A. GILLES
Salford