1984

Linguistics and indexing

David Crystal

Professor of Linguistic Science, University of Reading

In recent years, linguistics has developed a way of looking at language which may offer some insights to the indexer. Three main stages of inquiry are identified: observational, intuitional and evaluative. It is suggested that evaluative discussion of indexes is dependent on prior research at the observational and intuitional stages.

I must admit to a certain crisis of identity in addressing this topic. I am, first of all, a member of the large clan who regularly use indexes. Secondly, I am a member of the somewhat smaller clan who get cross when indexes are badly constructed, or omitted. I am, also, a member of the much smaller clan who insist that all the books they have written should have an index (including—or perhaps I should say, especially—those written for children). And I am a member of the tiny clan who insist on doing the index themselves. For the past year, I have been a member of the even tinier clan who have found themselves doing it for others.* And, since last year, I have been a member of perhaps the tiniest clan of all—those who subscribe to *The Indexer*. Given all these *personae* competing for attention, it is thus with some relief that, for the present paper, I find myself able to fall back on a category which distances itself equally from all—the minuscule clan of general linguists.

The 'restricted language' of indexes

To a linguist, an index is a member of the class of 'restricted languages'—a term used by the British theoretician J. R. Firth to identify those varieties of a language where the possibilities of novelty and creative variation are minimal or non-existent, and where all the usage possibilities can be expressed using a very small set of rules. All varieties of a language are restricted to some degree, for by definition a variety refers to a situationally constrained use of language; but the characteristics of the situations usually allow a great deal of flexibility, and the rules governing appropriate usage are mostly highly complex. Religious English, scientific English, legal English, business English . . . varieties such as these can be intuitively identified with ease, but it is far from easy to carry out a linguistic description of their salient or most appropriate linguistic features, for each subsumes a great deal of variability. By contrast, there exists a small number of highly restricted varieties where the task, on the face of it, seems much more straightforward, as with the written language of heraldic inscriptions, the spoken language of BBC weather reports from coastal stations—and the language of indexes.

Because indexing language seems so restricted in scope, it has not received much attention from linguists, who have been preoccupied with the more complex systems of expression encountered in such varieties as everyday conversation or written narrative. Indeed, I do not recall ever having seen an article in which the linguistic properties of indexes were examined. But in principle it could be done—and, I would argue, it should be done. If the precedents of other areas of linguistic analysis are anything to go by, several points of mutual interest to linguists and indexers would emerge. It is now well recognized that, as a result of the formal description of conversational English, it has proved possible to define in a more precise way the nature of the learning problems in fields as diverse as foreign-language teaching and speech therapy. (Not only have these applied fields benefited incidentally, but insights have been gained into the nature of language which have proved valuable to those concerned with theory.) Similarly, the linguistic study of indexing language may help to define the nature of the problems facing all those involved with indexes—both as writers and as users. But this is to look to the future. I do not know of anyone who has carried out such a description, and in the present paper all I can hope to do is discuss some of the factors which would have to be borne in mind by anyone attempting the task.

The 'grammar' of indexing language

As a restricted language, indexing has its own rules which it is the business of linguistics to make explicit and formalize. This is the main legacy of Noam Chomsky's approach to the subject. Chomsky has taught us the necessity of making explicit our intuitions about the language we unconsciously use, and shown us something of the difficulties we encounter when we begin that task. Anyone who is a native speaker of English 'knows' his

*The index to R. Quirk *et al, A Grammar of English* (Longman, 1985), whereof the suffering will be recounted in due course: 'Indexing a reference grammar'.

Abridged version of the address given to the Society of Indexers, Bristol, 8 July 1983.

language, in the sense that he is a fluent speaker, who can recognize a well-formed English sentence when he hears one, and can correct an ill-formed sentence. No reader of this paper would have any difficulty spotting the error in the sentence *The men is looking at a car*, for instance, and all would be able to correct it. On the other hand, for the thousands of native speakers of English who can carry out this task, there are few who could explain what it was they were doing, by using appropriate terminology (such as *plural subject, singular verb* or *subject-verb concord*), or who could relate the kind of error observed in this sentence to others found in apparently unrelated sentences (such as *He looked at themselves*). There is evidently a difference between 'knowing' a language and 'knowing about' a language: the former is tacit knowledge, whereas the latter is conscious, explicit knowledge. And when our tacit knowledge is made explicit in this way, and written down as elegantly and succinctly as possible, the result is known as a 'grammar'.

In a grammar of English, then, we expect to find all the rules which govern the construction of sentences in the language. There ought also to be a discussion of problem cases (usage controversies in particular) where the rules do not seem to work. Chomsky emphasized that a good grammar would specify *all* the rules governing *all* possible grammatical sentences in a language—all the possible permutations which mature native speakers have learned to recognize as part of acceptable English, and which constitute their linguistic 'competence'. By contrast, it would not be much concerned with inadequacies which occur when speakers actually put their language to use—the hesitations, false starts, rephrasings, and other imperfections that inevitably take place, which Chomsky summarized under the heading of 'performance'. In fact, to make sure that linguists' attention was not distracted by these performance limitations, but was focussed properly on the underlying competence of a speaker, he postulated an 'ideal speaker–hearer' as the object of study—a hypothetical being who was free of all the errors and limitations which flesh is heir to. A good grammar, he argued, ought to be able to explain everything that an ideal speaker–hearer could do with his language. And one thing more: a *really* good grammar would also have certain features built in which would, as it were, guarantee its efficiency by ensuring that the analysis, as far as possible, lived up to such criteria as comprehensiveness, simplicity and elegance. Linguistic theory, Chomsky felt, was not yet up to the task of making explicit the criteria which would guarantee the maximum adequacy of a grammar in this way (an 'evaluation procedure'), but this was an essential long-term goal of the subject.[1]

To what extent can this model be applied to the study of indexing language (which, for the sake of convenience, can be economically identified as *I* [= Indexese])? It should be noted that there are three steps in this way of proceeding. First, there has to be an observational step: samples of *I* usage have to be obtained from mature 'native writers' of *I* (i.e. indexers). Secondly, the intuitions of these native writers have to be tapped, to determine whether agreement exists over what counts as a well-formed and an ill-formed 'sentence' (i.e. entry) in *I*. In cases of disagreement ('usage controversies'), criteria need to be made available for deciding on problem cases. Thirdly, evaluative criteria need to be drawn up, to determine the grounds for asserting that different exemplars of *I* (different indexes) are 'better' or 'worse' than each other. An essential point to appreciate, in this way of thinking, is that each step is dependent on the previous one: we cannot evaluate the adequacy of different indexes (step 3) until there is agreement about what counts as a set of well-formed entries, and how to handle problem cases (step 2); and judgements about well-formedness, in turn, derive from the indexer's observational training and experience, which have led him to produce samples of *I*, whether as amateur or professional (step 1).

What counts as an indexer's 'competence'?

An important point about competence and performance must be made in relation to this first step. If we want to arrive at a satisfactory theory of *I* (that is, the competence of *I* writers and users), the initial samples of *I* must as far as possible be free of the extraneous performance limitations which are the bane of an indexer's existence—such as the arbitrarinesses imposed by publishers' budgets and insensitivities. The index samples ought to be seen as the products of an 'ideal writer–user', who does not have to be bothered by such constraints. Only in this way will it be possible to focus clearly on the question of what counts as an indexer's 'competence'. The issue of how one relates the 'ideal index' to the harsh realities of publishing has, of course, to be investigated (as part of a 'performance theory' of indexing); but until there is some measure of agreement over what the form of an ideal index is, any such studies and discussions are bound to remain *ad hoc* and unsatisfying. The first step, then, is to accumulate samples of *I* which the indexer is satisfied with. What may not be clear is how much data of this 'ideal' kind exists. There may actually be less in print than one would like, due to the pervasive nature of the above constraints. If this is the case, then it will be incumbent on indexers to provide the unconstrained samples as a basis for study. But do indexers who have produced a product to the best of their ability, which has been arbitrarily altered by others, routinely retain their originals?

It is my impression that, in general, discussion of *I* has paid insufficient attention to the problems attendant on step 1. In particular, the limitations of the 'native speaker' analogy need to be attended to. Firstly, there are plainly no 'native' indexers, as there are native (mother-

tongue) speakers of a language. One does not learn to index at one's mother's knee—though the expertise of many contributors to this journal makes me sometimes wonder! On the other hand, it has to be assumed that the gradual process of using and reflecting upon indexes will in due course produce a 'competence' of varying maturity, which professional training sharpens and defines. The ability of a group of indexers to recognize 'instinctively' a good or a bad index when they see one—to agree that it is good or bad, without necessarily being able to say why, or to agree on why—constitutes part of the evidence that such a competence exists. The fact that indexers will agree about certain *desiderata* for index use, in relation to the needs of the reader (see below), is a further indication.

But obtaining agreement amongst professional indexers is not the whole of the story: indeed, this evidence constitutes only a tiny part of the identity of *I*. Far more important is the need to obtain evidence from the other 'speaker' of *I*—the index user. And here we encounter a second problem with the 'native speaker' analogy. A native speaker is a native hearer, and *vice versa* (barring pathological cases); however, while the indexer is also an index user, the reverse situation does not obtain—most users of indexes have never tried to write one, and their skills may be negligible or non-existent. In a real sense, they do not 'speak the same language'. The indexer is thus part of a curious one-sided 'conversation': he constructs sentences in *I* whose intelligibility to the audience he is addressing is quite uncertain, and he is never (or, at least, hardly ever) told what degree of success has been achieved. Other indexers may tell him, of course, and occasionally a reviewer may mention the index; but systematic feedback from the general user is lacking. At a research level, therefore, it would seem crucial to obtain as much information as possible on the intelligibility of indexes to the general user. And here I do not mean obtaining general comments about the value of an index, such as might be obtained from a questionnaire, but actual video-recordings of people using indexes, so that one can observe how they proceed, and how they succeed. It ought to be possible to set up experimental situations, using an indexed book within which entries were systematically varied in different ways. For example, one could time the response rate for reaching a desired piece of information, or plot the strategies used in order to arrive at the information, and determine whether these corresponded to expectation. I am too much of a neophyte to know whether this is routinely done; but it is important research to do, if the intuitions of the 'native' index users are to be properly described.

Moreover, if the implications of the term 'native' are to be followed through, then it will be important to obtain evidence concerning the 'naturalistic' acquisition of indexing skills—such as the letter-by-letter preferences shown by 12–13-year-olds in a recent study.[2] Just as it is possible to show that very young children, before they are formally taught to read, have clear and consistent views about the nature of the reading process,[3] so it might be possible to show that children have developed some views about the nature of indexes before being formally introduced to them. The acquisitional perspective is, from a linguistic viewpoint, one of the most important factors to take into account when devising a linguistic theory—though I doubt whether indexers would wish to go so far as to claim, as Chomsky does for spoken language, that the relevant abilities are 'innate'!

Problems of evaluating indexes

By contrast, most of the debate about the rights and wrongs of indexing would be located in relation to step 3 in the above procedure. Step 3 is plainly in evidence: the distillations of experience which emerge in the form of training manuals, British Standards, and the like. These are attempts to identify evaluative criteria—measures which will guarantee indexing consistency and adequacy, in relation to topic treatment and formal presentation—and as one reads through back numbers of this journal, it is evident that it is in relation to these matters that most of the discussion takes place. But, as an outsider, I am struck by the discrepancy which seems to exist between the extremely positive statements several writers have made about the aims of indexing theory, and the extremely negative statements which have been made about the achievements of indexing practice. On the one hand, there are statements which seem to have achieved the status of axioms, such as 'The needs of the reader must be paramount in providing the right index for any book' and 'The right index for any book is that which will give readers the best possible data retrieval system'.[4] On the other hand, these have to be set against such comments as: 'Even today we know relatively little about the best form in which indexes are to be presented so as to be of optimal usefulness',[5] and the following remarks of Richard Hyman:[6]

> Textbooks do not agree even on specific mechanical procedures, e.g., whether *See also* references should come at the beginning or end of an entry. Even less consistent are their explanations not merely of *how* but of *what* to index. Neophytes are advised to index *everything*, but *only* everything crucial, significant or pertinent. These adjectives are left undefined, though used repeatedly . . . Experienced indexers will have developed their own concepts of 'pertinent' as distinguished from 'peripheral'. Beginners will seek authoritative guidance, only to be told that any explanation must depend on the context, and to be warned that over-indexing is a cardinal sin . . .

I cite Hyman's comments at some length because his concern with explanation ('for so definable a product,

explanations of its creation are surprisingly elusive') identifies the problem as falling clearly within the third step of the above procedure—what Chomsky would refer to as a grammar's 'explanatory adequacy'. And Hyman is by no means alone in his concern. Hans Wellisch, for example, makes a similar point about pertinence in a comment about the correspondence which followed his paper on the alphabetization of prepositions. He considers that 'the whole issue of "important" versus allegedly "unimportant" index words is totally spurious and irrational'; and he quotes Marvin Spevack, who also claims that 'it is well-nigh impossible, linguistically or otherwise, to arrive at a satisfactory distinction between "significant" and "insignificant" words'.[7]

Can such matters be resolved? In my view, the first move in this direction must be to understand why such problems arise in the first place. One explanation, using the linguistic model outlined above, could be this: that there has been too much attention devoted to the evaluative matters of step 3, without an adequate foundation in step 2 (or, for that matter, step 1). But, the argument goes, it is not possible to resolve step 3 issues without the foundation of the other steps. The focus of research ought therefore to be on ways of 'getting at' the information on which step 2 depends.

This information, it will be recalled, relates to indexers' judgements concerning the well- or ill-formedness of indexes—more precisely, of the set of entries which constitute an index. I do not know whether there are studies which take sub-sets of entries (or even whole indexes) and systematically obtain reliable data on acceptability, but if Robert Collison's comment is anything to go by, it would seem not: 'The time has surely come when we should commission the setting up of sample indexes in different fonts and in different sizes of type, so that we can study how best indexes may be presented'.[8] If this has not been done for issues of formal presentation, it is unlikely that it will have been done for the more abstract matters of topic treatment. But the operative words here are 'systematically' and 'reliable'. Simply juxtaposing two sets of entries, and impressionistically supporting one format rather than the other (as is often done when, say, the relative merits of word-by-word vs. letter-by-letter indexes are cited) is not enough. Efforts must be made to devise ways of focussing on one variable at a time. More important, methods have to be devised of asking informants to react to the variable at issue without begging the question by prejudging the terms of their response. If you present two sets of entries to informants, differing only in (say) the placement of prepositions, the set of responses you will get if the informant's attention is directed solely to the contrast between initial and end placement will not always be the same as those you will get if he does not have his attention focussed in this way. In an investigation where the informant is 'blind' to its purpose, the linguistic experience is that some quite unexpected responses can emerge.[9]

What would happen, for example, if several sets of entries were provided, to include (to continue the example) not only entries with initial- and end-placed prepositions, but also entries with no prepositions at all, or entries with prepositions put in places which no right-minded indexer would dream of? This last point may seem bizarre, but it is standard practice in linguistics to demonstrate the acceptability of a sentence by contrasting it with a clearly unacceptable one, or to establish what the facts are by presenting informants with a list of sentences, some of which are clearly acceptable (to the mind of the researcher), some of which are clearly unacceptable, and some of which are of unclear status. The linguist who thinks he knows what the facts of usage are in advance is naive indeed: unexpected responses abound, which sometimes cause him to question the very basis of his assumptions about a construction, and always force him to seek explanations (perhaps in terms of the dialect background of the informant). Will not the same situation obtain in the study of *I*? The equivalent of dialect difference certainly exists, for example, in the different house-styles of publishing houses, in American vs. British practice, and possibly in the different recommendations of training courses. But even within one 'dialect', I wonder what would happen if the following range of possibilities were presented to a group of informants:

> the alphabetization of prepositions in indexes
> alphabetization of prepositions in indexes, the
> of prepositions in indexes, the alphabetization
> prepositions in indexes, the alphabetization of
> in indexes, the alphabetization of prepositions
> indexes, the alphabetization of prepositions in
> alphabetization, prepositions, indexes
> alphabetization, prepositions, in indexes
> alphabetization, of prepositions, in indexes
> alphabetization, of preposition, in indexes, the

and of course there are various other possibilities, through deleting and transposing the constituents. If the informants were asked to rank-order these, for example, what would be the result? Would there be a systematic response, and would the *a priori* less desirable entries all turn out to be dismissed? The reader might care to carry out the task and then compare notes with a colleague or two. The important point, in such work, is to establish whether shared judgements of well-formedness exist among native writers of *I*. It must not be assumed in advance that these exist, or what they are, though in any project the researcher will of course have a hypothesis in mind.

Clarifying meanings: the linguist's possible role

The nature of hypotheses about *I* raises a final ques-

tion, and suggests a possibly fruitful area of interaction between indexers and linguists. It is evident that many of the hypotheses about *I* are highly complex, and will need to be broken down into their constituent propositions if they are to be tested. Hypotheses involving such notions as 'the best possible data retrieval system' or 'the needs of the reader' cannot be tested as they stand. Or consider the notions which require attention, in this respect, in the following comments on editing: 'finding the exactly right formula of words . . . to bring together all cognate material'; collapsing material into a single entry involves 'a matter of indexing judgment, taking into account the relative weight of the textual material, and the closeness of the subject matter'.[10] My concern is over what is meant by such notions as 'cognate', 'relative weight' and 'closeness of subject matter'. Or, to refer back to the discussion of prepositions, to what is meant by 'important', 'pertinent', and the like. I do not share the scepticism of the authorities quoted there, suggesting that these matters are beyond scientific enquiry—though, equally, I do not doubt the difficulties involved in studying them. What must not be forgotten is that, although *I* is a restricted variety of a language, its very status as a variety means that it will reflect certain properties of the language as a whole. The indexer, especially if he is a native speaker of the language, cannot help but bring his intuitions about semantic structure and frequency to bear on his task. The reader, for whom the index has been devised, will also use his intuitions in this way. The more that is discovered about a language's semantic properties, therefore, the more it should be possible to clarify what is meant by the above notions. It is this point of contact between linguistics and indexing which I find especially intriguing. Investigating the interrelationship will be a challenging task; but at the end of the road is a linguistic theory of indexing language, and that is a goal worth pursuing.

References

1. Lyons, J. *Chomsky*. 2nd ed. London: Fontana, 1977.
2. Hartley, J., Davies, L. and Burnhill, P. Alphabetization in indexes: experimental studies. *The Indexer* **12** (3) April 1981, 149–53.
3. Ferreiro, E. and Teberosky, A. *Literacy before schooling*. London: Heinemann, 1982.
4. Hall, B. M. Getting the index right: roles and responsibilities. *The Indexer* **13** (3) April 1983, 166.
5. Wellisch, H. H. The alphabetization of prepositions in indexes. *The Indexer* **12** (2) October 1980, 91.
6. Hyman, R. J. Indexes for analysis and diagnosis. *The Indexer* **13** (3) April 1983, 177.
7. Wellisch, H. H. Letter to the Editor. *The Indexer* **13** (1) April 1982, 48.
8. Collison, R. L. The future of indexes and indexing. *The Indexer* **12** (4) October 1981, 172.
9. Quirk, R. and Svartvik, J. *Investigating linguistic acceptability*. The Hague: Mouton, 1967.
10. Hall, *op. cit.* 167.