Indexing a reference grammar

David Crystal

The article gives an account of the procedure used in indexing the reference grammar by R. Quirk, et al., *A comprehensive grammar of the English language* (London: Longman, 1985). It outlines the main characteristics of a reference grammar, and the kinds of problem facing the indexer of such works. A chronology of the indexing operation is given, which involved the compilation of a preliminary index on cards, and subsequent discussion with the authors, leading to the production of the final version by word processor. The main features of the index are identified, along with the kinds of modification which resulted from the joint discussion. The importance of co-operation between indexer, authors, and publisher is stressed.

*A comprehensive grammar of the English language*, by Randolph Quirk, Sidney Greenbaum, Geoffrey Leech, and Jan Svartvik, appeared in 1985, published by Longman, weighing in at 2.5 kg (x + 1,779 pp.). It is the latest in a series of reference grammars written by this team, the first of which was published in 1972, *A grammar of contemporary English* (1,120 pp.). That is a lot of grammar, by any criterion, and it requires a lot of index to go with it. But indexing a reference grammar brings to light some interesting problems.

**Reference grammars**

What is a reference grammar, first of all? The most succinct characterization is to say that it does for the grammar of a language what a dictionary does for the vocabulary. If you want to know the meaning of a word, or its spelling, or some aspect of its usage, you look it up in a 'reference lexicon'—in other words, a dictionary. Correspondingly, if you want to know about the meaning of a grammatical construction, or its form, or how it is used, you look it up in a 'reference grammar'.

There have been several such works written during the 20th century—the most famous being Otto Jespersen's *A modern English grammar on historical principles*, in seven volumes (London & Copenhagen, 1909-49). Other reference grammarians include E. Kruisinga, G. Scheurweghs, and R. W. Zandvoort, and the discerning reader will note the distinct lack of Anglo-Saxon names. At least the present grammar has one pure-blooded English name in the list (but not two—Quirk is Manx).

The analogy with the dictionary is instructive. The series of grammars which this team have produced may be likened to the various 'levels' of the *Oxford English dictionary*, or any comparable lexicographical venture. There are 'shorter', 'concise', and other editions of a dictionary; and so it can be with a grammar. The 1972 book, considered enormous in its decade, now has to be seen as the 'shorter' edition. Quirk, et al 1985 is the equivalent to the ' unabridged'.

The analogy with the dictionary breaks down in one crucial respect, however. To use a dictionary, all you have to know is the order of letters in the alphabet and how to spell. Armed with this knowledge, you can look anything up. Vocabulary is not linguistically organized in an alphabetical way, of course, but the separation of semantically related items (such as having *aunt* at one end of the book and *uncle* at the other) does not usually pose major problems for most users. In a reference grammar, however, the biggest problem is how to organize the information to enable readers to find what they are looking for. Alphabetical order will not help here. It makes no sense to organize a grammar along alphabetical lines (though it has been tried)—page 1, abbreviations; page 2, abstract nouns; page 3, active voice; and so on. On this basis, 'active voice' might be on page 3, and 'passive voice' on page 333; and there would be similar silly splits between 'definite' and 'indefinite' articles, 'past', 'present', and 'future' tenses, and so on. The whole point of grammar is that meaning is expressed through systems or networks of contrasts—achieved through altering a word-ending, adding words, omitting words, and permuting them in various ways—and all the contrasts that belong to one grammatical system need to be dealt with at the same place. The grammar has to be organized more on the lines of a thesaurus than a dictionary, with chapters devoted to different domains of grammar, interconnecting and overlapping in many ways.

And so we find, in this grammar, a sequence of chapters which begins by outlining the way verbs are used in English, then nouns and determiners (*the*, *my*, *this*, etc.), then pronouns and numerals, then adjectives and adverbs, and so on until one reaches the more complex aspects of syntax, which take in the structure of clauses, sentences, and units larger than the sentence (such as the paragraph). The difficulty, of course, is how to find one's way about, without having to read the whole thing first.

There are four ways of attacking this problem. The first is the equivalent of knowing alphabetical order: a general knowledge of the structure of the language, obtained from previous reading or courses. This will
provide a sense of where material ought to be located. The second is to give the reader a survey chapter, in which the conceptual organization of the work is explained (as in a thesaurus) and the main technical terms introduced: this is done in Chapters 1 and 2. The third is to ensure adequate cross-referencing, to enable the reader who enters the grammar at a given point to discover the superordinate and subordinate notions related to his point of enquiry, and thus to develop a sense of the shape of the whole of the grammatical wood, as well as the identity of its constituent trees. And the fourth, of course, is to ensure that the work has a good index—because even with the first three under your belt, there is still a major problem in finding the facts about the usage of an individual construction. Take the issue of *I shall* vs. *I will*, for example. Will this be discussed under *shall? will?* tenses? future tenses? auxiliary verbs? modal verbs? All these identifiers are accurate. Every area of grammar poses such problems of retrieval; and for the indexer, a critical question is 'How redundant should the index be?'

**General considerations**

The authors had accepted that the index to the 1972 book was inadequate for the needs of the reader. This was unfortunately so. Empirical linguistic research tends to be highly detailed and meticulous, with careful attention being paid to the way a form is used in all parts of the language. It is not easy to generalize about such things as *will* and *shall*, or the use of *whom* vs. *who, or where only* should be placed in a sentence. Every relevant instance of the use of such items may need to be scrutinized, before reaching a conclusion. It needs to be possible to extract from the grammar every piece of discussion about *whom, or shall, or only*, regardless of where it occurs. Similarly, a researcher might be interested in a general factor governing the patterns of usage in the language—for instance, whether the pattern is British or American, formal or informal, from speech or from writing. It should be possible to track down every instance of an American English usage in a reference grammar, every instance of an informal usage, and so on. But in the 1972 book, this was not possible. Although American English uses are scattered throughout the book, there is but a single reference in the index to 'American English', referring to where this notion is introduced in a general way in the opening chapter. Formality is similarly referenced only in Chapter 1; and generally, one would have to conclude that the work is under-indexed. It has 26 pages of index out of 1120 pages of text: just over 2%.

My main principle, for the new work, was to provide a much more detailed index, in an attempt to anticipate the kinds of linguistic enquiry which might one day be made by users of this book. If people were going to be interested in American vs. British English usage, then it was important that *all* substantial references were itemized. Similarly, all relevant discussion of individual words and constructions of grammatical importance (such as *but, do, shall, of, only*) would be given a comprehensive indexing, along with any relevant sub-classification (there are, for example, seven different grammatical uses of *only*). The main consequence, of course, was a considerable increase in the size of the index. The entries for American and British English, for instance, brought together some 500 references; the entries on formal and informal had over 600. The result was an index of 6265 main entries (i.e. unindented head-words, with or without textual references), taking up 113 pages: 6% of the book.

**Entry analysis**

However, the description in terms of entries is misleading, when a closer analysis is undertaken. Leaving aside cross-references, there are two kinds of main entry in this index: general notions ('coordinator', 'copula', etc.), and specific lexical items—individual affixes, words, or phrases (*dare, data, de-, deal with*, etc.) (see below, for the significance of this distinction). The former constituted 985 main entries, with an additional 2161 sub-entries; the latter constituted 4536 main entries, with an additional 904 sub-entries. The different main-entry/sub-entry ratios reflect the different kinds of information in a grammar: the high ratio of the former (2.19 per main entry), reflecting the need to bring together under one heading related grammatical notions, contrasts with the low ratio for the latter (0.2 per main entry), reflecting the idiosyncratic behaviour of individual lexical items. A relatively high number of *see also* references (2101) also reflected the 'network' property of a grammar. Main entries and sub-entries combined totalled 8586. Only 37% of these were general notions, but (in a 20-page sample) these accounted for 53% of the chapter/section references made.

The index also contains a large number of straight cross-references (1188). Some of these are simply abbreviation expansions (cop → copula), there being a considerable number of abbreviations in a reference grammar. Some arise from the decision to group a set of entries in a single place (all categories of 'name', for example, are placed under the one heading). But most are due to one of the main aims of a reference grammar, which is to be theoretically eclectic—to include references to the different ways of analysing grammatical patterns encountered in the literature on English grammar, and thus to the associated terminology. (Linguistics, ironically, is a field which has produced an extraordinary amount of terminological variation in recent years.*) For example, 'subordinate' clauses are often known as 'dependent' or 'included' clauses; a construction with elements omitted might be referred to

*See, for example, my A dictionary of linguistics and phonetics (Oxford: Blackwell, 2nd edn., 1985).
as 'elliptical', 'reduced', 'truncated' or 'abbreviated'. The index has got to allow for this variability. There is no way of predicting which term might be used as an entry-point to the index, given the diverse theoretical backgrounds of the possible users. A substantial amount of cross-referencing seemed to be the only way to get around this problem.

Chronology of the operation

My experience of indexing this book provides an excellent example of the kind of co-operation between author(s), indexer, and publisher often commended in the pages of this journal. I came onto the scene in December 1982, following the first stage of revision of the text, after each of the four authors had taken into account the comments of the others, and of one additional reader. The authors were planning to produce a third draft, incorporating the comments of other readers, and this in turn would be the basis of a final revision, due to take place at a summer vacation 'write-in' in London in 1983. (The authors had written their respective sections at their home bases of London, Lancaster, Lund and Milwaukee, and this would be their first and only opportunity to draw all the loose ends together.) The question arose as to when I should be sent copy on which to base the index: it was suggested that the best time would be following the summer revision, once the final text had been agreed, but I felt that a preliminary index would be a helpful tool to have available during the write-in, given the amount of cross-referencing in the body of the text which was envisaged, and the fact that the authors had been working in isolation from each other, and were thus likely to have produced material which would need to be made consistent. I therefore began work using the second draft, as it became available from the authors. I was sent the second (overview) chapter first, which was helpful, as it gave me a sense of my major headings; and subsequent chapters came through in random order over a 4-month period.

The preliminary indexing brought to light, as expected, a large number of points of inconsistency and overlap, and raised several queries of a more substantial kind. The original intention was for me to send any comments on a chapter directly to the author concerned; but it quickly became apparent that the kinds of query which the indexing was raising were of equal interest to all four. The authors therefore incorporated me into their correspondence system, whereby any questions raised by one of them were automatically circulated to all. I was thus able to see any changes in thinking during the last year of the text revision, and they, of course, were able to see the kinds of queries which the indexing of a particular chapter was raising at any time. This meant, for example, that a point of terminological inconsistency noticed early on in the indexing could be taken account of during the revision of the later chapters.

For instance, an inconsistent use of the symbol 'A' was found in Chapter 2, being sometimes used for 'adjunct' and sometimes for 'adverbial'. Alerting the authors to this kind of problem undoubtedly saved time at the write-in.

In all, just under 4,000 points arose in the preliminary indexing, which largely involved the following kinds of issue:

1. Absent, unclear, inconsistent, or wrong cross-references to other sections of the book (a persistent question was which section to set up as the 'basic' point of cross-reference, when a topic was treated in several chapters).
2. Terminological inconsistency between sections or chapters. Some points were minor (e.g. the choice between 'participle phrase' and 'participial phrase'); others were potentially serious, raising questions of analysis and definition. For instance, a set of verbs which included such items as say and tell were variously called verbs of 'utterance', 'communication', 'saying', 'speaking', 'discourse', 'declaration', 'narrating', 'reporting', and 'speech act'. It was not at all obvious whether the authors were referring to the same verbs by these labels, and it proved necessary to sort out this (and similar problems) at the write-in.
3. Several terms were located which had not been previously defined, especially in the introductory chapters.
4. The usual problems of consistency of typography (e.g. whether to italicize an item, whether to hyphenate) came to light.
5. The book contained a number of summary tables (e.g. irregular verbs, prepositions), which did not always correspond to the text.
6. An unexpectedly large number of usage labels, characterizing a style or effect, were discovered (e.g. 'abruptness', 'admonitory', 'archaic', 'artistic', 'awkward')—over 250 in all. Several seemed to be used synonymously. A separate list was therefore made of these, so that they could be considered together, and thus reduced to more manageable proportions.
7. Several observations were made about the balance of notes to main text, the relationship between main and subheadings, and so on. Occasionally, my linguistic background enabled me to make suggestions of substance, but these were kept to a minimum, given the advanced stage of the book's preparation.

The authors' third draft (some 3,000 pages of A4, double-spaced) had been a considerable expansion of their second (in the light of comments received), and they expected to reduce the size considerably at the write-in. I took my cue from this policy, and made the preliminary index much larger than I expected ultimately to need, and built in maximal redundancy (e.g. I indexed 'noun phrase' under both 'noun' and 'phrase'). In this way, I was able to obtain a better sense of the balance of infor-
mation in an entry, and also discovered something of the
authors' preferences for look-up (information which is
rarely available to the indexer).

I had asked for guidelines for the index, but the
authors had no firm views—except that the index was to
be more detailed than before. I proposed various basic
points, and these were discussed and agreed. The alpha-
betical arrangement was to be word-by-word, to ensure
that semantically related items would be as close together
as possible. References were to be to chapters and
sections, not pages. References would be numerically
explicit, with no use of or . Passim would be used only
when a topic was discussed throughout the whole of a
chapter. A reference to a section would automatically
direct the reader to any notes accompanying the section;
however, if an item turned up in the notes alone, this
would be separately designated (e.g. 2.16n), and no
further distinction would be made between notes, if more
than one was attached to a section. Bibliographical
references would not be indexed, nor would the names of
sources of data (e.g. examples taken from named books
or newspapers). Italics would be used to indicate major
references within long entries. The authors were typing
their final version on grid pages provided by the pub-
lisher; I would be given a grid also.

One index or two?

I decided to divide the indexing data into two, for the
preliminary work, reflecting the two levels on which
linguistic studies operate. On the one hand, there was the
conceptual apparatus of English grammar; on the other,
there was the set of lexical items which the grammar
spent a great deal of space discussing. Non-linguists
there was the set of lexical items which the grammar
conceptual apparatus of English grammar; on the other,

One index or two?

I decided to divide the indexing data into two, for the
preliminary work, reflecting the two levels on which
linguistic studies operate. On the one hand, there was the
conceptual apparatus of English grammar; on the other,
there was the set of lexical items which the grammar
spent a great deal of space discussing. Non-linguists
there was the set of lexical items which the grammar
conceptual apparatus of English grammar; on the other,

One index or two?

I decided to divide the indexing data into two, for the
preliminary work, reflecting the two levels on which
linguistic studies operate. On the one hand, there was the
conceptual apparatus of English grammar; on the other,
there was the set of lexical items which the grammar
spent a great deal of space discussing. Non-linguists
there was the set of lexical items which the grammar
conceptual apparatus of English grammar; on the other,
Cards or disk?

I bought my word processor—an AES Alphaplus 12—just after the write-in. The preliminary index had been done on cards, as was evidenced by my torn fingers. But there had been problems in using cards. In particular, the random presentation of chapters had been a real problem. Receiving Chapter 13 first, then Chapter 4, then 9, makes for problems in deciding how to allocate the space on a 6" x 4" card, and the result in many cases was, to put it politely, a graphic mess. The amount of revision at the write-in had been too great to enable me to use the same cards (many section numberings had been revised, and there were many deletions and additions of text), so it was necessary to start from scratch (though, of course, most of the time-consuming decision-making about the conceptual levels involved could be taken over from the preliminary index without change). The final versions would again be arriving in random order, and there were likely to be several last-moment revisions of points of detail which would affect the index. It seemed wise to switch to the word processor, which could handle the erratic arrival of material without difficulty.

It certainly was a wise decision. Apart from the pleasure of working with 'clean' copy at any one time, I estimate that the saving in time was about a third—and it would have been more, if I had been more skilled in the use of my machine at the time. There were a few problems, of course. The biggest was to decide how much information to include in a memorized page. If I included too much, the page took a relative age to update for each additional entry, especially as the disk got full. On the other hand, if I kept the pages too short, I was continually having to find the relevant page-name in my page reference guide, in order to recall it, and this took time, as a typical page-name would be BRE-BRIT. As the index grew, disk changing also became a nuisance. I tried to cut corners by scanning pages for entries in alphabetical order, but this did not really save any time: I

Part of a page from the index to A comprehensive grammar of the English language.
made a short time-and-motion comparison at one point, and found that the time I spent repeatedly scanning a page, finding all entries beginning with *ba*, and the like, was more than the time it took to take the entries in sequence, look up the name of a disk page, recall it, and update it. And even using a distinctive red underlining, Murphy's law would operate: having just entered ten items beginning with *con-* and updated the page, my eye would fall on the eleventh *con-* item, gaily waving from mid-paragraph.

In the end, I used 312 disk pages, which printed out 378 A4 pages, double-spaced, using a column width of 45 characters. I had no problems with accidental erasing, and the like, because the Alphaplus system comes with a valuable recovery program which enables one to retrieve any page which may have been inadvertently erased. The system also comes with two disk drives, which enabled me to have immediate access to about 60 pages at any one time, without having to change a disk.

**The home straight**

Immediately after the write-in, we agreed presentation conventions with the publisher's sub-editor, e.g. which symbols I should use to represent arrows and the swung dash (neither available on my processor), how I should distinguish between dash and hyphen, and so on. I produced a sample, which the sub-editor revised, and we agreed on a final format. I was able to keep in direct touch with the sub-editor during the last few months, and his role was crucial in ensuring that any index changes relating to the last-minute revisions submitted by the authors (some 600, in all) were accurately incorporated into the typescript being sent for typesetting (or, later still, into the proofs).

The final (sic, see above) version of the index was circulated to the authors in January 1984. I pointed out that the index was now at the stage that the main text had been in the previous summer, and that it should be revised in a similarly thorough way. In particular, bearing in mind the need for 'user-friendliness', I asked the authors, when reading through, to bear six points in mind:

(a) to check all head words, to agree their potential relevance for the future user;
(b) to check that the head words made immediate sense, and did not include any ambiguities;
(c) to check the grouping of terms under a single heading, to ensure that my intuition coincided with theirs;
(d) to check the italicizing of major references, in case I had overused, or underused, this convention;
(e) to check that they approved of the few special 'features' I had introduced (e.g. the compilations on names, verbs, and terminological issues).

I received, in due course, some 500 points to take into account—mainly suggestions about italicization of major entries, but with some suggestions about entry regrouping and extra cross-referencing. A few further entry deletions were made, and the authors got their own back by pointing out various inconsistencies in the way I had used my own conventions.

The authors were pleased with the index. As one said, having received the final version, 'Now we can find out what is wrong with the book!' They graciously acknowledged the role of the operation by adding a reference to the index on the title page. The task took 18 months. 'Suffixing a name by an obelisk', say the authors on the last page of the grammar, in the appendix on punctuation, 'indicates that the person is dead'. It was the last item to be indexed in the book, and it provided, almost, an accurate designation of this indexer's physical and mental state.

David Crystal is Honorary Professor of Linguistics at the University College of North Wales, Bangor; author of several books on English usage; broadcaster; editor (currently of Linguistics Abstracts); and indexer.

**Specialized vocabularies**

We have received an updated edition of the *ILO thesaurus: labour, employment and training terminology* (Geneva: International Labour Office, 1985, 463 pp. 30 cm. bibl.) An introduction explains how to use the thesaurus. That is followed by three alphabetical KWOC indexes in English, French, and Spanish, and a systematic display, with notes on the use of the descriptors. Terminology added or changed since the last edition (1978) reflects economic and social changes recorded in the literature of the past few years. New terms include Flexible retirement, Youth unemployment, Female-headed households, Return immigration; and among revised terms Work attitude replaces Work ethic. It is hoped that in addition to its primary function of recording and retrieving information in the ILO Library's LABORDOC data base, the thesaurus will also be used by other agencies and so facilitate the exchange of information in its field.

From the United Nations Translation Division have come two terminology bulletins: *Units of currency* (1984. Bull. no. 329, superseding no. 325) and *Names of countries and adjectives of nationality* (New York, 1985. Bull. no. 333) a trilingual list, in English, French, and Spanish, of states which are members of the United Nations or its specialized agencies, or parties to the Statute of the International Court of Justice, together with two provisional lists of terms in the same languages, the first list relating to the UN Development Programme, *Terminology UNDP*, 1985, and the second to the work of UNICEF, *UNICEF: provisional list of terms*, 1982.

M.P.