1987 Krishnamurthy, R. 1987. The Process of Compilation.
In Sinclair, J.M. (ed) Looking Up: An account of the COBUILD Project in lexical computing. London: Collins ELT

Chapter Three: The Process of Compilation by Ramesh Krishnamurthy

Although this paper is intended to be a factual account of the process of the compilation of the Cobuild Dictionary as it actually happened, and not a discussion of theories about how best to set about compiling a dictionary, it may be useful to look briefly at the following passages, in order to place our methods and decisions in some kind of context.

Zgusta [1971;p.223] gives us a theoretical model for compilation when he breaks down `the work to be done by the lexicographer' into four main tasks:`
1. the collection of material;
2. the selection of entries;
3. the construction of entries;
4. the arrangement of the entries.'

For the details of the tasks involved, we can compare the following two accounts, widely separated in historical terms:

In 1791, Boswell [1906 edn;p.110] describes how Johnson put together his dictionary in the following way: `The words, partly taken from other dictionaries, and partly supplied by himself, having first been written down with spaces left between them, he delivered in writing their etymologies, definitions, and various significations. The authorities were copied from the books themselves, in which he had marked the passages with a black-lead pencil...'

Kipfer [1984;p.1] outlines the practicalities of current techniques of dictionary-making: `Most users believe that the lexicographer simply sits down and "writes" a dictionary. This is far from true! The dictionary editors conduct a reading program, excerpting quotations (citations) from...written...and...spoken sources. These citations ...may be stored in computer...The editors take all the slips for the word and divide them into the different senses, then use them to write the actual definitions...A dictionary is a record of actual usage.'

Apart from the reference to `etymologies' in the quotation from Boswell, which is a feature not relevant to synchronic learners' dictionaries such as the Cobuild dictionary, the main elements of the process of compilation have remained constant since his time. Kipfer, with her reference to a rudimentary computer-held corpus of data, comes closer to describing the process of compilation as it actually occurred at Cobuild.

For the purposes of this paper, I propose to consider the process of compilation in two main sections: resources and process. Some illustrations of output are given in the

Appendix.

1.RESOURCES:

The Cobuild lexicographer had at her or his disposal the following major resources: the corpus, policy papers, works of reference, expert linguists in the Birmingham University English Department, native-speaker informants, and the computer.

1.1.THE CORPUS:

In the Cobuild Dictionary project, the first of the tasks listed by Zgusta, `the collection of material', remained largely out of the hands of the lexicographers throughout, because a concordanced corpus of 7.3m words (6m words of written text and 1.3m words of transcribed speech) was extracted from the Birmingham Collection of English Text and made available to lexicographers in the form of microfiches and `hard copy' (printed pages). A more detailed description of the corpus and its development will be found in chapter 1.

The concordances, representing the `authorities' in Boswell's terms or the `citations' in the quote from Kipfer, were the primary tool of lexicographers on the Cobuild project and where the evidence of the corpus was clear, its influence on the dictionary entry was always decisive. Not only were the concordances unique to Cobuild, but the project was the first to work directly from such a large body of evidence stored in this form.

In the later stages of compiling and for both of the on-line editing phases, fiches of concordances were available from a corpus of about 20m words for forms with fewer than 50 occurrences in the 7.3m word corpus (for example, for the form `bereaved', there were only 6 concordance lines in the 7.3m corpus, but a further 25 lines came from the larger corpus).

In addition, where further guidance was required on the way in which a particular lexical item was commonly used in EFL coursebooks, etc, access to the TEFL Side Corpus was provided (for example, there were 11 lines for the word `adjective' in the 7.3m, 23 more in the 20m, and 65 in the TEFL Side Corpus; the comparable figures for the form `illustrating' were 7, 17, and 26).

Frequently occurring words present a great problem for lexicographers in trying to analyse the evidence. As Murray [1888;p.xi] says, `...with the larger articles, as those on AT, BY, BUT, BE, BEAR, BREAK...the mere study of the result, arranged in some degree of order, gives little idea of the toil and difficulties encountered in bringing into this condition what was at first a shapeless mass of many thousand quotations.'

At Cobuild, this problem was greatly reduced by the creation of a sampling program that provided any specified proportion of concordance lines on an `every nth line' basis for any frequently occurring items for which the total concordances were too numerous to be easily analysed by lexicographers.

The standard concordance line consisted of 132 characters with the keyword in the middle, and the lines were normally presented in alphabetical order of the first character of the word following the keyword. A set of codes at the left of each line indicated whether the line was from the spoken or written part of the corpus, the nationality of the author and the country of publication, and the specific text or transcript from which the line was taken. Thus the lexicographer was given some guidance as to the language mode, register, and possible regionality of each line.

However, for words such as conjunctions, discourse organizers, or signal words that required a longer context for proper analysis or appropriate exemplification, a program was written to produce a selection of longer concordances from a 1m word sub-corpus. For the word `as' therefore, the longer concordances produced the following long examples which eventually found their way into the dictionary text: `Napalm should be banned, as should the development, production, and stockpiling of all chemical weapons.' `He was totally unprepared, as is the way with American he-men, for anything that could not be settled with a fist or a gun.' `He thinks he would like to teach, but as his subjects are Greek and Moral Philosophy he's not likely to find a job.' `The mother (as if she didn't have enough to do already!) has to remember to pay some attention to her husband.'

For some indication of lexical words that might fall into this group, see Eugene Winter's list of `Vocabulary 3' items: Proposed Lexical Items of Connection [1978;p.97] which includes such words as `basis, case, cause, compare, condition, distinction, matter, instance, point, reason, way, etc'. Examples selected for the final dictionary text again show the advantage of being able to have available the longer concordance lines: for `reason', `Public pressure is towards more street lighting rather than less: the reason is, of course, that people feel safer in well-lit streets.'; and for `way', `I've been given six months to do the job. A week one way or the other will make no real difference.' The longer examples preserve the naturalness of the type of discourse in which these `Lexical Items of Connection' operate.

The same program that produced the longer concordance lines also offered the facility for alphabetically ordering the concordance lines by the character string to the left of the keyword, so that leftward collocations such as preceding prepositions, determiners, quantifiers, modifiers, or other lexical items were more easily identified.

This was very useful in the analysis of a word like `time', for example, which often figures at the rightmost end of phrases and prepositional groups. There are 9481 lines for `time' in the 7.3m word corpus (which is in itself a daunting number of lines for any lexicographer to scrutinize), and 1594 in the 1m sub-corpus. By selecting just 300 of these, the following leftward collocations emerged: 16 lines for `at the same time', 14 for `all the time', 13 for `at the time', 13 for `a long time', 11 for `for the first time', and so on. Phrases such as `once upon a time' and `one...at a time' were also easier to spot, even when there were only one or two lines for them.

Finally, concordances were produced for second elements in compound words where a hyphen provided the element-break marker, so that entries for combining forms such as `-toothed', `-looking', etc could be compiled.

The optical scanning process, by which much of the corpus was input to the computer, produced a fair number of misreadings (e.g. 200 concordance lines for `going' were found under the form `gong', lines for `well' under `weli' and for `the' under `thc'). Proof-reading has been carried out subsequently.

The lexicographer was still left facing the problem of identifying multi-word items in which the words did not appear consecutively in the text, such as phrases with considerable lexical variation (e.g. `shake someone by the hand, shake someone's hand, shake hands with someone, etc') or phrases with `open slots' into which any of a set of lexical items can be inserted (e.g. `a fisherman's dream, a footballer's dream, a politician's dream, etc'), and, of course, separable phrasal verbs (e.g. `put off' as in lines such as `some deliberate ploy to put the reader off', `They decided to put the whole thing off until the following day' etc.)

1.2.POLICY PAPERS:

Lexicographic policy had been developed from 1981 onwards by experimentation and trial compilation. A set of policy papers was gradually established which encapsulated the Cobuild house style.

In the latter half of 1983, a thorough review of policy was undertaken and the results incorporated in a revised set of policy papers. These aimed to deal with the problems raised by the old policy papers, to clarify policy where divergent interpretations or ambiguities had become apparent, to bring the compilation of the dictionary database more into line with the requirements of the dictionary itself, and to fill gaps in policy and make explicit policies that had developed over the previous two years.

These new policy papers formed the basis for all subsequent compilation and policies remained fairly fixed from this point on, apart from a few minor modifications (e.g. the addition of SWH as an abbreviation for `somewhere' in phrases that required a following adjunct of place or direction) and a few architectural changes in the dictionary database which were effected by program (for example two-word compounds such as `driving licence' and `green belt', previously held as phrases became headwords in their own right, and phrasal verbs such as `drive off' and `swear by' held as `phrases' became categories and were tagged in a look-up table that enabled the phrasal verbs to be output as sub-entries).

Eventually, in 1985-6, many of the entries that had been compiled under the old policy were recompiled so that the whole of the database reflected a consistent policy.

The policy papers issued in late 1983 consisted of 32 papers, each dealing with a particular topic:

1. The headword list and selection of senses:
The main criterion for inclusion or exclusion of headwords and senses was the strength of the corpus evidence.

2. The slip:
The type of information to be recorded on the computer input slips and the number of characters available in each field was explained as well as the mechanics for arranging and numbering the slips in a sequence (see Appendix: 2 for slip formats).

3. Headwords, lozenge words, and second alpha order:
The criteria for treating lexical items as headwords, derived words (lozenge words in Cobuild terms because of the symbol that was generated to indicate such words in the dictionary), or phrasal verbs (which were to appear in a subsidiary alphabetical order at the end of headword categories, hence `second alpha order') were set out. There was a list of acceptable suffixes (e.g. -ly, -ness, etc) for items treated as derived words, and rules for when such derived words were to be treated as headwords.

4. Categorization:
The categorization of a word was primarily based on semantics. Differences of word-class and syntax were used as the basis for creating sub-categories. Ordering of categories depended mainly on frequency of occurrence in the corpus data, with concrete or literal senses preceding abstract or metaphorical ones, while maintaining some cohesion in the semantic flow through the entry.

5. Definitions:

Definition style was fairly rigorously laid down, with the main emphases on accuracy and clarity, while avoiding difficult vocabulary and constructions. The main principle of definition consisted of providing a genus word and then supplying sufficient differentiae to distinguish the headword from its near-synonyms. For further details see Chapter 6.

6. Usage restrictions that affect definitions:
Where there were severe syntax restrictions on a word or sense, for example a verb that was only used in the passive, the commonest syntactic structure was defined rather than the headword form.

7. Examples:
Examples for a word or sense were selected in order to show typical usage, involving syntactic patterns, regular collocations, and appropriate contexts, rather than to clarify or extend the definition. They were taken from the concordances wherever possible, but slight modification was permissible.

8. Phrases:
Three basic types of phrases were isolated: fixed phrases (e.g. `by and large' and `once in a while'), syntactic phrases (e.g. `on ONE's own' and `give SB the pip'), and lexical phrases (e.g. `now and then, now and again' and `a good deal, a great deal'). Multi-word items functioning as prepositions (e.g. `in relation to') or subordinating conjunctions (e.g. `on the grounds that') were also covered.

Guidance was given on how to define phrases, at which element to place them, and where to place them within an entry. Polysemous phrases and phrases with pragmatic import were discussed. The distinction between phrases, compounds, and collocational or syntactic patterns was indicated.

9. General syntax:
The system of syntax notation was specially developed to allow lexicographers to record not only word-classes and paradigmatic syntactic variations, but also individual syntagmatic sequences. This was achieved by the use of a considerable number of labels or `primitives' linked by a small set of logical or relational symbols or `connectors' according to fixed conventions.

The labels included the usual VB for verb, N for noun, ADJ for adjective, etc, but also QUAL-BRD for broad qualifiers (including demonstratives, possessives, modifying adjectives, and qualifying adjuncts or clauses), NEG-BRD for broad negatives (including strict negatives such as `no' and `not', adverbs like `seldom', and some uses of items such as `few' and `only'), etc.

The connectors included such symbols as `+' (followed by or having in its environment), `~' (usually used in a particular tense, functioning as a particular syntactic unit, etc), `/' (realized lexically as), `( )' (optional feature or element), `< >' (composed of or functioning as), and so on.

Thus, typical notations for items included such strings of labels and connectors as `N-AN//S' (noun always with the determiner `a' or `an' and never used in the plural), V+PREP/FOR (intransitive verb always followed by the preposition `for'), `PHR-VB<V-OD+ADV>' (phrasal verb composed of a transitive verb and an adverb), and `ADJ-EAP+NEG-BRD' (qualitative adjective used in both attributive and predicative positions and always having a broad negative in its environment).

To ensure maximum flexibility, wherever the syntax notation system or indeed any other aspect of policy was felt to be inadequate to cope with a particular piece of information, the letters XP were written on the appropriate part of the computer input slip and a note written within square brackets wherever on the slip there was sufficient space.

10. Nouns:
The syntactic coding for a noun involved compulsory recording of information regarding number, use with determiners, and countability. Syntactic requirements (e.g. always followed by the preposition `of' or a `that' clause) were also a matter for compulsory recording.

Additional syntactic information could be recorded by using labels indicating that it functions as an itemizer or quantifier, refers to a container or to the amount contained, is a proper noun, is used in titles or forms of address, etc.

Guidance was also given on the treatment of plural nouns, the nominal use of adjectives and participles, and nouns that typically occur in a particular clause position or that can be used in the singular or plural with no semantic difference and no real denotation of number.

11. Compounds:
The distinction was drawn between nominal compounds, phrases, and nouns used to modify other nouns. The criteria included the transparency of the combination, given the individual meaning of the elements, and whether a clear paradigm operated at both slots. The commonest orthographic form of a compound (i.e. one word, two words, or hyphenated) was always indicated.

12. Adjectives:
The primary sub-divisions of adjectives was into qualitative adjectives, colour adjectives, and classifying adjectives. The positions they could occur in (i.e. premodifying, predicative, and postnominal) were an essential part of their syntax notation.

Adjectives commonly used as the head of verbless clauses (e.g.`nervous and trembling, he opened the letter'), after `it is', or before a clause introduced by `that, who, if, etc' were given a separate notation.

The use of adjectives in the comparative and superlative, as nouns, or as object complements (e.g. `He cut the bread thick'), and the recording of adjectival inflections were also dealt with.

13. Adverbs:
Only four main types of adverb were recorded for the database: adverbs that could modify an adjective, another adverb, or a verb; adverbs of degree; sentence adverbs (or disjuncts); and broad negative adverbs (e.g. `hardly' in `He was hardly able to speak.'). Further information could be recorded by using the general syntax notation (e.g. to note that `asunder' always follows the verb it modifies) or adding a special note, for example to indicate aspective adverbs such as `technically' and `financially'.

14. Verbs:
Verbs probably have the widest range of syntactic variability in their environment and this was reflected in the policy paper on verbs. Transitivity was obviously one of the key features to record, and the basic symbols used were V for intransitive verbs (or, to be more precise, verbs which cannot take a lexical object, but may take a clause, infinitive, or participle as object or be followed by a complement) and V-OD for transitive verbs.

In addition to these, labels could be added to specify ditransitives, ergatives, reflexives, performatives, reporting verbs, modals, auxiliaries, verbs that have impersonal subjects (such as `it', `there' or `what'), cognate objects, etc.

Problem areas such as delexical verbs (e.g. `heave' in `heave a sigh') and compound verbs (e.g.`bulk buy') were also discussed. Features of the syntactic environment, for example infinitives, participles, complements, `that' clauses, adjuncts, etc could be indicated using the general syntax labels. Usage restrictions such as for verbs never used or always used in a particular tense or mood were also coded, as well as syntactic patterns. Inflections were dealt with in paper 21.

15. Phrasal verbs:
Distinguishing phrasal verbs from literal combinations (e.g. `refer to' and `face forward') and completive combinations where the adverb simply added a sense of `completely, `repeatedly', or `intensively' (e.g. `deal out (cards)') was a problem, and depended on such criteria as whether the same verb occurred in the same meaning independently or with a range of other particles or adverbial phrases, and whether the particle occurred in the same meaning with other verbs.

Phrasal verbs were held as categories in their own right at the end of the main entry for the headword. Restrictions on the separability of the particle from the verb or on the possibility of passivization were also noted. Infrequent combinations such as `throw overboard' and `throw open' were treated as phrases.

16. Participles:
Participles were also a problem area, and the paper discussed how to deal with the adjectival use of a present or past participle, when to treat it as a main entry, a derived word, or merely as an example within a verb category. The distinction between adjectival past participles and passive verbs was discussed. Guidance was also given for dealing with participles used as combining forms or nouns.

17. Combining forms:
Lexical items (e.g. `-toothed' and `-looking'), often participles, that were used productively in the formation of compounds were treated as main entries. Different information was recorded at each element, depending on which was the productive one.

18. Prepositions:
Most of the common, single-word prepositions were compiled as a set in a separate special operation. Prepositional phrases (e.g. `on account of') and prepositions derived from other lexical items (e.g. `considering') were treated as main entries.

19. Conjunctions:
All the co-ordinating conjunctions and most of the single-word subordinating conjunctions were compiled separately as a set, so only phrases functioning as conjunctions (e.g. `on the grounds that') and conjunctions derived from other lexical items (e.g. `considering' when followed by `that') were discussed.

20. Interjections:
The label for interjections could be used for categories, phrases, or individual examples.

21. Inflections:
Inflections for verbs, nouns, adjectives, and adverbs were discussed. For each word-class, the regular inflections were prescribed and needed no notation. Common but irregular inflections and variations between American and British English forms (e.g. `traveling' and `travelling') were coded. Highly irregular inflections were given in full (e.g. for `go') and any alternatives were noted (e.g. `learned' and `learnt').

Inflectional information was only given for the first category of each word-class in an entry and was assumed to be the same for all other categories having the same word-class unless new information was supplied. If it was, it only applied to the category for which it was given and had to be repeated if it applied to more than one category.

The last element of a compound was assumed to inflect, and the information given depended on the policy for its word-class. Any irregularities were given in full. Phrasal verbs were not given inflections unless there was any change from the inflections given in the main entry.

Irregular inflected forms were treated as main entries (or categories in the case of homographs such as `bore') and cross-referred to the uninflected form.

22. Pronunciation:
Although lexicographers were not responsible for supplying the phonetics at the compilation stage, we had to indicate any change in pronunciation within an entry at the relevant category (e.g. for `contest', at the first verb category after the noun categories or vice versa). For abbreviations, truncated forms, acronyms, etc, a note was required about whether the form was pronounced as individual letters, as a word, as the expanded form, etc. Alternative pronunciations were noted.

23. Synonyms, antonyms, and superordinates:
We were given guidance on the selection and recording of these semantically related items for a category, phrase, derived word, or for an individual example. Database cross-references were made to the individual elements of any multi-word items entered.

A crude notion of core meaning elements in definitions was established and synonyms were expected where possible to share, and antonyms to be antonymous to, most of the core elements, to be the same word-class or phrase type, and close in register. Synonyms would therefore be near-synonyms of each other, but antonyms might be antonymous to different core elements.

Given any necessary syntactic adjustments, synonyms for an example were to be substitutable for the headword in it without much change in register. Antonyms were to be substitutable for the headword plus a negative.

The principle of superordinateness was founded on an equally crude model of lexis in which, for example, `scottie, terrier, dog, canine, and mammal' belong to adjacent levels in a hierarchy, and therefore `terrier' is the superordinate for `scottie', `dog' for `terrier', and so on.

Only one superordinate was allowed for each category or example and had to be the same word-class as the headword, could not also appear as a synonym, but could also appear as a semantic field label.

24. Field:
The term `field' referred to the semantic field of a word. Items entered under this heading were to be nouns or nominal

groups, for ease of retrieval and future use. The number of items was restricted by the number of characters available on the computer input slip. Field labels were chosen from the next level up in the hierarchical notion of lexis used for superordinateness.

Thus for `bad-tempered', field labels would probably include `people, character, behaviour, and irritation'. Despite the freedom allowed to us in the selection of field labels, early reviews of the database revealed that we were being remarkably consistent in our choices.

25. Style:
Style labels indicated how a word was used, for example whether its use was more prominent in a particular geographical region, in speech or in writing, in formal, informal, or intimate situations or contexts.

Some labels indicated the attitude of the user (e.g. rude, pejorative, offensive, euphemistic, approving, or humorous), others the likely type of user (e.g. children, or a particular group or profession, such as criminals, journalists, sportsmen, computer staff, doctors, lawyers, or poets). Archaic, old-fashioned, current or neologistic uses were labelled, as were trademarks, proverbs, cliches, metaphoric usages, and non-standard usages. Multiple labels could indicate simultaneous or alternative features.

26. Pragmatics:
During the early stages of compiling, the need to record the pragmatic force or effect of a word or utterance became evident, and a range of methods was developed to allow this. Explicit performative verbs (e.g. `object, agree, promise, and predict'), concealed performatives (e.g. `Please accept my sincere apologies' = `I apologize sincerely'), and implied performatives (e.g. `I didn't mean to be rude' = `I apologize') were identified and labelled.

Where the pragmatic function of an utterance was not so evident, but the pragmatic effect on the listener or reader could be assessed, this was noted (e.g. the persuasive effect of `I'd be really grateful if you could help me'). Other such effects included dissuasion, deception, encouragement, and bribery.

Utterances indicating the relationship of a speaker or writer to their discourse, e.g. commenting on it, structuring it, or moving it from one idea or subject to the next were also noted. These were often signalled in written text by occurring at the beginning of a sentence or between commas or dashes (e.g. `actually' in `Actually, I don't think that's quite right' politely signals the correction about to be made). Phrases such as `by the way', `so to speak', or `and so on', and closed-set turns such as `Thanks' and `Do' were also noted for their pragmatic effect.

27. Collocates:
The definition of regular or significant collocates was `lexical items occurring within five words either way of the headword with a greater frequency than the law of averages would lead you to expect'. The importance of collocation was stressed and we were asked to record these items in dictionary examples. Function words such as prepositions, determiners, pronouns and auxiliaries were ignored, but delexical verbs were significant. Collocation was established only on the basis of corpus evidence.

28. + Box:
The + box on the computer input slip was used to record collocates, related words (e.g. night and nocturnal), items other than the headword in phrases, compounds, and combining forms, and elements of multi-word items recorded as synonyms, antonyms and superordinates. It could also be used to generate a cross-reference to any item whose compiler might benefit from information recorded at the source entry.

29. Cross-references:
This summarized the information on cross-references given in other papers and gave a few more cases where they were needed. For example, internal database cross-references were automatically generated by any item entered as a synonym, antonym, or superordinate, and any item in the `+ box' (see 28).

Surface or Text cross-references were intended to appear in the dictionary text to indicate to the user where a particular word, phrase, or sense was treated. Where known, the particular category of the target word was given. A less common synonym or synonymic multi-word item could be defined simply as `another word for...' or `another expression for...'. Stylistic or regional restrictions could be specified (e.g. `a formal word for...', `an American word for...'). Headwords also appearing as derived words in another entry had to cross-refer to them. Polysemous phrases in different categories of an entry cross-referred to one another. Semantic confusables (e.g. `foetus' and `embryo'), irregular inflections, and alternative spellings and forms required cross-references, as did phrasal verbs and their nominalizations (e.g. `take over' and `takeover').

30. Alternative spellings and forms:
The full entry for a word appeared at its commonest spelling or form and a cross-reference was placed at the other spelling(s) or form(s). Concordance lines selected as examples retained the original spelling or form. No examples were given at the less common spelling(s) or form(s).

Details were given for alternatives of the `despatch/dispatch' type, for American spellings (e.g. `color' for `colour'), for British spellings (e.g. `gaol' for `jail'), for polysyllabic verbs ending in `-ise' and `-ize' (with a few exceptions, `-ize' was the preferred main form) and initial capitals (merely exemplified if infrequent, but treated in a separate category where always capitalized in a particular sense). Derived words and phrasal verbs did not show any alternative spellings. Items written as one word, two words, or hyphenated, were compiled at their commonest form and the alternatives mentioned if common. A compound and one of its elements (e.g. `scuttle' and `coal scuttle', `thermos flask' and `thermos') or a shortened and unshortened form (e.g. `bra' and `brassiere') were treated as synonyms where appropriate.

31. Abbreviations and acronyms:
No distinction was made between abbreviations and acronyms. Most abbreviations were compiled in a separate operation, so we needed only to attend to extremely common ones (e.g. `TV') or homographs of items being compiled (e.g. `salt' and `SALT'). The presence or absence of full stops was ignored. Pronunciation was indicated (i.e. whether an abbreviation was pronounced as letters, as a word, or as the expanded form). Definitions were supplied if the abbreviation was the commonest form, otherwise only a cross-reference was needed. Inflections (e.g. `pp' as the plural of `p' meaning page) and syntax were stated as for any other item and polysemous abbreviations were categorized as any other headword.

32. Illustrations:
At the compilation stage, the dictionary was expected to have illustrations, so wherever an item was felt to need one, it was so marked.

1.3.WORKS OF REFERENCE:

These included most of the reference tools that any linguistic project in the English language might need to consult on occasion.

At our disposal were the major international dictionaries such as Oxford English Dictionary and its Supplements, various editions of Webster, a range of dictionaries produced by Oxford University Press and Longman for native-speakers and EFL learners, and of course, as funding publishers, a large selection of Collins dictionaries, including bilingual ones (as many of the lexicographers had a foreign language at degree level) and especially the Collins English Dictionary.

We also had recourse on occasions to various grammars of English, especially Quirk et al, as well as to thesauri, usage books, and EFL coursebooks. There was a reasonable selection of encyclopedias, specialist dictionaries such as the Penguin series, volumes on sport, wildlife, and other technical subjects. We had numerous works, including journals, articles and papers, covering the theoretical aspects of lexicography. Any gaps in our in-house collection were usually readily filled by the University Library.

All of these were, of course, secondary tools, and consulted only after inspection and analysis of the corpus data.

1.4.EXPERT LINGUISTS IN THE BIRMINGHAM UNIVERSITY ENGLISH DEPARTMENT:

The project was part of English Language Research within the English Department of the University, and consequently we could call on assistance from colleagues in the department. They conducted seminars in various aspects of syntax, lexis, and discourse, and lexicographers attended specialist seminars at ELR. Various members of ELR also gave us guidance on specific types of lexical item such as discourse organizers, disjuncts, etc. Numerous visitors to the project and outside speakers gave us valuable insights into different areas of language, both theoretical and practical.

1.5.NATIVE-SPEAKER INFORMANTS:

Under this heading I would include not only the informal and frequent consultations between fellow-lexicographers on the Cobuild project, during which personal intuitions could be tested and confirmed or rejected, but also the various forms of language input that we all encounter in daily life. Newspapers, radio and television programmes, advertising material, overheard conversations, etc, often helped to fill the occasional gaps in the corpus data.

Information acquired in this way always remained a minor and usually secondary input, but was extremely useful in the area of neologisms or new senses of existing items (e.g. 'massage' in the expression `massaging statistics'), or for the more colloquial and spoken lexical items. For a period of time, a `neologisms file' was maintained and regularly updated to provide references for words that had acquired recent importance.

However, queries on some items could not be resolved even from data of this kind: for example, when trying to ascertain whether the term `L plates' (for the signs required to be displayed on a car being driven by a learner driver in Britain) was normally written with a hyphen (L-plates) or not, a thorough scrutiny of The Highway Code and Advice for Learner Drivers (the two official HMSO publications that might have yielded some evidence) failed to provide even a single occurrence of the term, although the displaying of `L plates' is a legal requirement !

1.6.THE COMPUTER:

As already mentioned in section 1.1, the computer allowed us to gain access to the data in the corpora in various different ways.

As compiling progressed and data was input to the dictionary database, programs were written to provide screen or printed output of database entries. Thus increasing evidence became available from words already compiled by fellow-lexicographers.

Eventually, the computer was used to automatically generate, for any word about to be compiled, printouts of cross-references from other related items together with the relevant parts of the database entries for those items (for example, for the other elements in a phrase, for collocates, for synonyms, etc), and these were issued to lexicographers with each batch of words to be compiled and the relevant concordances.

Facilities for interrogating the dictionary database using almost any item (e.g. a syntax pattern, a particular lexical item entered as a synonym, etc) were made available from an early stage and these facilities were refined and extended as the need arose. They proved invaluable in standardising syntax notation during the extraction of the dictionary from the dictionary database.

Information could also be retrieved from the compilation monitor database, such as the headword list itself, or details about which lexicographer had compiled a word and when, how many concordance lines there were for a headword in the 7.3m word corpus, and in which of the three dictionaries and three TEFL sources (from which the headword framework had been created) the headword occurred.

The computer was used to overhaul the dictionary database prior to extraction of the dictionary, and became the central tool of the project during the editing stages, when most of the work was carried out on-line.
Fuller details of the role of the computer in the Cobuild project will be found in Chapter 2.

2.PROCESS:

Sections 2.1 and 2.2 roughly cover the second area of the lexicographer's task as defined by Zgusta, namely 'the selection of entries'.

2.1.ALLOCATION OF ITEMS TO LEXICOGRAPHERS:

The creation of a headword list for the dictionary database and its use as the basis for the compilation monitor database (for details see Chapter 2) meant that words could be allocated to lexicographers with a great degree of flexibility and efficiency.

Batches of words could be selected from an alphabetic sequence, or on the basis of semantic similarity, or because they belonged to a closed or restricted set.

Words could be selected by the project controller or by the lexicographer, and the risks of omission of a headword or duplication of effort were vastly reduced.

2.2.CONCORDANCES: (`First catch your hare, hares, haring and hared'):

For each item allocated to lexicographers, concordances were issued to them or could be selected by them from the `hard copy library'. The first task of the lexicographer was therefore to ensure that concordances for all the relevant forms of the headword being compiled were to hand. For example, the singular and plural forms for nouns; base, comparative, and superlative forms for adjectives; all tense and participle forms for verbs.

This was especially important for irregular words such as child/children, buy/bought, etc. Forms ending in apostrophe and s were separately concordanced, as the computer notion of keyword was a simple `string of characters between spaces'. An initial capital letter was not a problem as such occurrences were concordanced along with initial lower case ones. Nominal compounds presented a further problem, as they could appear in three separate sets of concordances (i.e. as one word, hyphenated, or as two words). Alternative spellings such as words ending in -ise and -ize were obviously concordanced separately, but the lexicographer had to be constantly on the alert to recognize the possibility of variant forms and to check the concordances for evidence of those forms as well.

Before compiling an abbreviation or truncated form, or a headword that the lexicographer knew had an abbreviation or truncated form, it was necessary to check two or three sets of concordances in order to ascertain which form had the highest frequency. This was in order to ensure that the item was given the fullest treatment at its commonest form, and cross-references placed at the less frequent ones. Occasionally, for example at the entries for `MP' and `Member of Parliament', both forms received full treatment because both occurred with very high and almost equal frequency.

If no concordance evidence was available for a headword, in most cases the word was rejected for compilation, and this was recorded in the compilation monitor database. Occasionally, if the lexicographic team agreed that an item should nevertheless be compiled, for example if it appeared in another EFL dictionary or was a very colloquial usage that just happened never to have been used in the corpus texts, this was done, but the item was usually marked in the dictionary database as being held for the database only. Such items were not extracted for inclusion in the dictionary text. The same procedure was sometimes carried out on rarer senses of a headword.

Neologisms that were felt to be likely to remain in use for some time were compiled and the word was inserted into the compilation monitor database at the same time as the entry for it was keyed into the dictionary database.

2.3.CONCORDANCES: ANALYSIS

Section 2.3 begins the third phase of a lexicographer's task in Zgustan terms, i.e. `the construction of entries', but in fact Zgusta fails to specify what is in fact surely the central core of the lexicographer's art, the analysis of the material collected.

In checking that the concordances for all the possible forms of a word had been gathered together, a comparison of the relative frequency of the various forms gave an early indication of prominent word-classes, the potential sub-types of word-class. For example, a noun with no -s form concordances might indicate its uncountability. A verb with mainly -ing or -ed form evidence might indicate that its verbality had declined and its main use was as an adjective (e.g. there were no lines for `encrust' or `encrusts', one for `encrusting' used as an adjective, and 45 for `encrusted', only two of which could be construed as active past tenses of the verb, the rest being adjectival uses) or noun (e.g. 11 lines for `backsliding' as a noun and none for any form of the verb), etc.

In the early stages of compiling, coloured felt-tip pens were used to mark concordance lines in which the word (in whatever form) was used with a particular area of meaning. This was in order to facilitate subsequent grouping of examples drawn from concordances for different forms into the same semantic category, and also the ordering of the categories of an entry into frequency order. Later on, as lexicographers got used to working with concordances, this technique was largely dispensed with.

Collocational evidence was of great usefulness in an analysis of the corpus data. The concordance lines were arranged in alphabetical order of the first character after the space following the keyword. This meant that some features of the behaviour of a lexical item in text became immediately apparent.

A frequently occurring following

preposition probably indicated a syntactic requirement or pattern for the word or a particular sense of the word (e.g. 70 of the 96 lines for the form `refer' were immediately followed by `to'), many lines for the same following noun might suggest a nominal compound (e.g. 5 lines out of 19 for `staple' in the 7.3m corpus were followed by `diet' and 12 lines out of 62 in the larger corpus). Verbal forms often followed by nouns might, on closer inspection, give evidence of a typical object or type of object for a transitive verb. Pronominal items occurring in the following slot might indicate a phrase or pragmatic utterance (e.g. 50 lines for `mind' followed by `you' indicated the use of the expression to introduce an afterthought, modification, or warning).
Reflexives might suggest a reflexive verb.

Collocates occurring in other positions with respect to the keyword were often a little more difficult to recognize, but lexicographers soon became adept at registering these as well.

The distinction between a regular collocation for a particular word or sense and the more rigid pattern of a set phrase was sometimes very hard to make, because there is a continuum between them.

Collocates often helped to indicate semantic categories.

2.4.CONCORDANCES: SELECTION OF EXAMPLES

After an initial survey of all the concordances, the next task was the selection of examples. A more detailed discussion of the principles behind the selection of examples will be found in Chapter 7. The first example selected for any word or sense of a word was intended to show typical usage in terms of syntactic behaviour and collocation. Subsequent examples registered syntactic patterns, further collocations, etc.

Wherever possible, examples selected from the concordances were not altered in any way. However, some modifications were sometimes necessary. For example, a very long sentence might contain a great deal of material that was felt to be extraneous, irrelevant or confusing to a learner, and might be shortened to focus attention on the keyword and essential syntactic and collocational features.

Concordances from the spoken section of the corpus might contain false starts, revised sentence structures, hesitations, repetitions, etc which could be omitted without affecting the integrity of the example (e.g. in the concordances for `bit', there is a line `er no you feel a bit dizzy afterwards yes feel a bit and you have to take it easy for a bit you lie', which might be amended to `You feel a bit dizzy afterwards and have to take it easy').

Sometimes, an author may have inserted a word of a very different register from the keyword in order to create a particular stylistic effect. If the word could be replaced by one more compatible with the register of the keyword, this may have been done (e.g. in the concordance line for `vermin': `In Scotland, feral cats are treated as vermin, and so poisoned.', `feral' was changed to `wild' for the dictionary example). If the keyword was used twice within a sentence, but used with different meanings, it might be appropriate to delete one of the occurrences (e.g. `for a bit' in the example given a few lines above).

If any such changes were made to the actual concordance line, this was noted on the example slip. Some examples were felt to be of interest for future research but did not warrant inclusion in the dictionary, and these were marked accordingly, with notes on the reasons for holding them.

Only on very rare occasions, when the corpus evidence was exceptionally limited or idiosyncratic in some way, did lexicographers actually make up entire examples.

2.5.CATEGORIZATION:

Section 2.5 commences the third task ascribed to lexicographers by Zgusta: `the construction of entries'. The practical criteria involved in the process of categorization are more fully discussed in Chapter 4.

Once the examples had been selected for a headword, they were grouped together in categories on the basis of semantics and sub-categories were created where differences of word-class or syntax required them.
This often involved drawing the distinction between syntactic requirements and syntactic patterns (cf. policy paper 9 in section 1.2).

The ordering of categories depended on frequency of occurrence and precedence of concrete over abstract senses, with due regard for the maintenance of semantic flow within the entry (cf. policy paper 4 in section 1.2).

2.6.DEFINITIONS:

Murray [1888;p.vi] declares : `A Dictionary...is not a Cyclopaedia: the Cyclopaedia describes things; the Dictionary explains words, and deals with the description of things only so far as it is necessary in order to fix the exact significations and uses of words.'

Once the examples for an entry had been grouped into categories and sub-categories, and these had been ordered, definitions had to be written for each category and sub-category. Cobuild definition style aimed at accuracy and clarity, and the structural requirements were a genus word and sufficient differentiae to distinguish the headword from any near-synonyms. Esoteric vocabulary, highly idiomatic expressions and complex constructions were avoided. For further details see policy paper 5 in section 1.2 and Chapter 6.

2.7.ADDITIONAL INFORMATION:

By the time the lexicographer had reached this stage in compiling a word, much of the lexical (collocational) and syntactic environment had already been noted on the example slips. If it had not been done already at the time of categorization, the essential syntactic requirements for each category and sub-category had to be established, with the aid of the examples within it.

The compiler had now to check that, for each category and sub-category, any relevant information concerning the inflections, pronunciation, style, and pragmatics had been entered. A superordinate had to be supplied, and synonyms, antonyms and field labels sought. If an illustration was felt to be desirable, the pink slip was appropriately marked (see the relevant sections of 1.2 for details).

All that remained was for the ordered categories and sub-categories to be numbered according to a fairly basic decimal notation system, and for each slip within the entry to be numbered consecutively. Various procedures were available for subsequent insertion or removal of slips before keyboarding.

2.8.COMPILATION PHASES AT COBUILD:

At Cobuild, the overall process of compilation took place in four major steps:

2.8.1.COMPILING ON COMPUTER INPUT SLIPS:

Lexicographers analysed the corpus evidence, mainly from the 7.3m word corpus, and compiled entries on slips of paper that were specially designed to hold the information in a format suitable for computer input to the dictionary database. The entries were reviewed by senior colleagues and revised where necessary after due consultation with the compilers. Various groups of words, for example abbreviations, particles, auxiliaries, and modals were compiled separately in special operations.

2.8.2.INPUT TO DATABASE AND OTHER COMPUTER PROCESSES:

The entries were keyboarded from the slips into the dictionary database as compiling progressed. Later, computer programs were run over the database to achieve a greater degree of standardisation of entries, to clean up certain recurrent errors, etc. Some manual editing of the database was also carried out. Trial editing to produce dictionary text from the database was carried out and numerous typeset samples prepared and reviewed. The preliminary dictionary text was then extracted from the database by computer programs, which also inserted the relevant typeface codes.

2.8.3.ON-LINE EDITING OF DICTIONARY TEXT:

By this stage, it was necessary to introduce considerations relating to the desired shape and style of the actual dictionary entries. Each entry was to be introduced by a full list of the forms of the lemma involved. These were generated by programs, which however often proved inadequate to deal with the idiosyncrasies of the English lexicon, and a great deal of manual correction was required by lexicographers.

The database concept of `category' and `sub-category' was superseded by the `dictionary as prose' units of paragraph and sub-paragraph (see chapters 4 & 6). This entailed a major change in policy as regards the ordering of the information within an entry. The semantic ordering of categories in the database was found to lead to a considerable fragmentation of the structure of a dictionary entry as, for example, a verb category was followed by its related noun and then by another verb category and another noun, etc. Hence, for the dictionary, it was decided to reorder the categories on a syntactic basis. Again, computer programs were used to do this automatically, but only partially succeeded.

However, to accommodate cases where it was felt unnecessary to create a separate paragraph to illustrate the less frequent use of a word in the same meaning but in a different word-class, a new strategy was developed. This involved the generation of a new symbol for the printed text ( ) followed by an indication of the change of use: `used as a noun', `used as an adjective', etc. The same strategy was adopted for minor changes in the semantic application of a word in a particular sense, as for example an adjective that commonly described people but could also describe their behaviour, or a noun that referred to a container but could also refer to its contents.

Another major change involved the shift of emphasis from the traditional dictionary concept of `definition of meaning' to the principle of `explanation of usage', which considerably altered the style of language and structures used. See Chapter 6 for further details.

Information about syntax and synonyms, antonyms, and superordinates was to be placed in an extra column to the right of the main dictionary text, to highlight the information and to avoid breaking up the flow of prose.

Phrases were grouped together after the main senses of a word, except where a phrase was felt to draw specifically and only on a particular sense. Polysemous phrases were grouped together in a paragraph, as it was realised that a learner could not be expected to know which senses of a word gave rise to phrases identical in form.

During this on-line editing phase, as the fiches for the larger corpus were now available, we attempted to replace any made-up or greatly amended corpus examples with real extracts from the corpus. The larger corpus also gave us valuable new insights into many words, which changed some entries radically from the database ones. For example, the metaphoric use of the word `graveyard', (as in `Elections are the graveyard of the political prophet') signalled by the following preposition `of', was represented by only two lines out of 20 in the 7.3m word corpus, and therefore had not been included in the database entry, but gained an additional 9 out of 65 in the 20m word one, and was admitted to the dictionary entry itself.

Senior colleagues, as in the database compilation phase, reviewed the edited dictionary text and it was revised accordingly. The phonetic symbols to be inserted into the text to indicate pronunciation were automatically input to the appropriate part of each entry during this phase.

2.8.4.SECOND ON-LINE EDIT AND FINAL CHECKS ON DICTIONARY TEXT:

The resulting text was read and commented on by outside experts who had not been involved in the compiling. These comments were incorporated into the dictionary text if they were supported by the evidence, again by on-line editing.
Style checks helped to increase consistency of presentation within the dictionary.
Checks were made to eliminate spelling errors and to increase consistency in spelling where variants existed. A cross-reference check was instituted to ensure that dictionary users were not sent on fruitless hunts. A sensitivity check was also carried out to minimise undue or inadvertent offence.
Some proofreading corrections were also carried out on-line.

CONCLUSION:

The procedures described in this paper show a blending of traditional lexicography with modern technology. Much of the time and energy was spent in coping with unexpected events, because the methodology was quite untested. Progress was slow initially and accelerated as areas of the work stabilised, so that the final stage of extracting and editing the dictionary from the database was done in little over one year. We think we can say with Murray [1888;p.xi] : `Our own attempts lay no claim to perfection; but they represent the most that could be done in the time and with the data at our command.'

APPENDIX :

The appendix is intended to display the different formats in which information about a word was held at different stages of the Cobuild dictionary project.

1.CONCORDANCES FROM 7.3m WORD CORPUS AND LARGER (c.20m WORD) CORPUS: the format of concordances showing alphabetic ordering by first character to the right of keyword and codes identifying spoken or written source, nationality of author, and place of publication.

2.SLIPS: the format and typical information contained on the computer input slips (for details of actual input procedures, etc, see Chapter 2).

3.DICTIONARY DATABASE: the format of printed output from the database using the `dbpe' program (for details see Chapter 2).

4.MACHINE-EXTRACTED DICTIONARY TEXT: the format of printed output of computer files containing such text, with typesetting codes.

5.PAGE PROOF OF DICTIONARY TEXT: the format of the final dictionary page at the proof-reading stage.

REFERENCES:

1. Manual of Lexicography by Ladislav Zgusta;
Published by Mouton, 1971.

2. The Life of Samuel Johnson LLD by James Boswell Esq Vol 1;
Published by J M Dent in the Everyman's Library series, 1906.

3. Workbook on Lexicography by Barbara Ann Kipfer;
Published by the University of Exeter as Vol 8 of Exeter Linguistic Studies ed. R.R.K.Hartmann, 1984.

4. Preface to `A New English Dictionary on Historical Principles, Volume I A and B', edited by James A.H. Murray;
Published at the Clarendon Press, Oxford, 1888.

5. `A look at the role of certain words in information structure'

by Eugene Winter; in KP Jones and V Horsnell (eds),
Informatics 3, ASLIB, London 1978, pp 85-97.