Archibal Michiels and Jacques Noël

THE PRO'S AND CON'S OF A CONTROLLED DEFINING VOCABULARY IN A LEARNER'S DICTIONARY

In order to assess and improve a dictionary, we need to free ourselves from the limitations of entry-bound look-up. This is possible only if

(a) the dictionary is available in machine-readable form;

(b) its information categories are sufficiently formalized and formatted;

(c) software is available that allows us to formulate our queries simply and directly and to interrogate our data-base interactively.

We shall take up these three points in turn.[1] Our main preoccupation will be to compare two dictionary files, one using a free defining vocabulary and the other a controlled vocabulary, THE (NEW) COLLINS CONCISE ENGLISH DICTIONARY (CCED) and THE LONGMAN DICTIONARY OF CONTEMPORARY ENGLISH (LDOCE). These two dictionaries are of roughly the same size and their computer files have been made available to us by the publishers under research contracts. As non-native users of these dictionaries we tend to espouse the learner's perspective (to what extent does the LDOCE defining vocabulary represent 'core' vocabulary and at what syntactic price is lexical simplicity bought?), but we are also interested in the possibilities offered by a controlled defining vocabulary for the retrieval of semantic relationships within dictionary definitions.

We shall concentrate here on the relationship between an instrument and the process it is typically used in or for. This issue should be of interest not only to lexicographers, but also to all those who wish to develop tools for automatic or semi-automatic access to the information contained in dictionary data-bases.

In order to be fully exploitable by computer, a dictionary file ought to be much more than just machine-readable: its various information fields should be sufficiently formalized and formatted to allow access to individual categories of information. In this respect, LDOCE is unique in providing a whole range of clearly demarcated information categories (see Fig. 1).

Especially for the purposes of this paper it would have been desirable for the data-base itself to allow for the demarcation of examples and definitions. Unfortunately, even in LDOCE, complex typographical criteria would have had to be used to make that distinction; thus we had to leave this job for future research. As far as the CCED categories are concerned, the only information field that stands out clearly is the headword. All the information pertaining to a given headword is presented en bloc. Our computer scientist therefore had to write a program based on an algorithm using typographical information in the dictionary file in order to isolate the fields represented in Fig. 2.

Fig. 1 LDOCE information categories

1. HEADWORD
2. PRONUNCN
3. POSINFLE (Part of Speech + inflexions)
4. ENTRYCOD (grammatical and other codes for whole entry)
5. DEFINCOD (codes applying to a single definition)
6. DEFINTXT (text of the definition; also includes examples)
7. USAGNOT (usage note)
8. ENTRYTYP (includes homograph number and information on number of words and respective POS in multiple word entries)

Fig. 2 CCED information categories

1. HEADWORD (headword number and identification)
2. POSINFLE
3. DEFINTXT

Each definition, with its accompanying examples, was isolated. Various types of meta-linguistic information, such as etymology, were discarded.

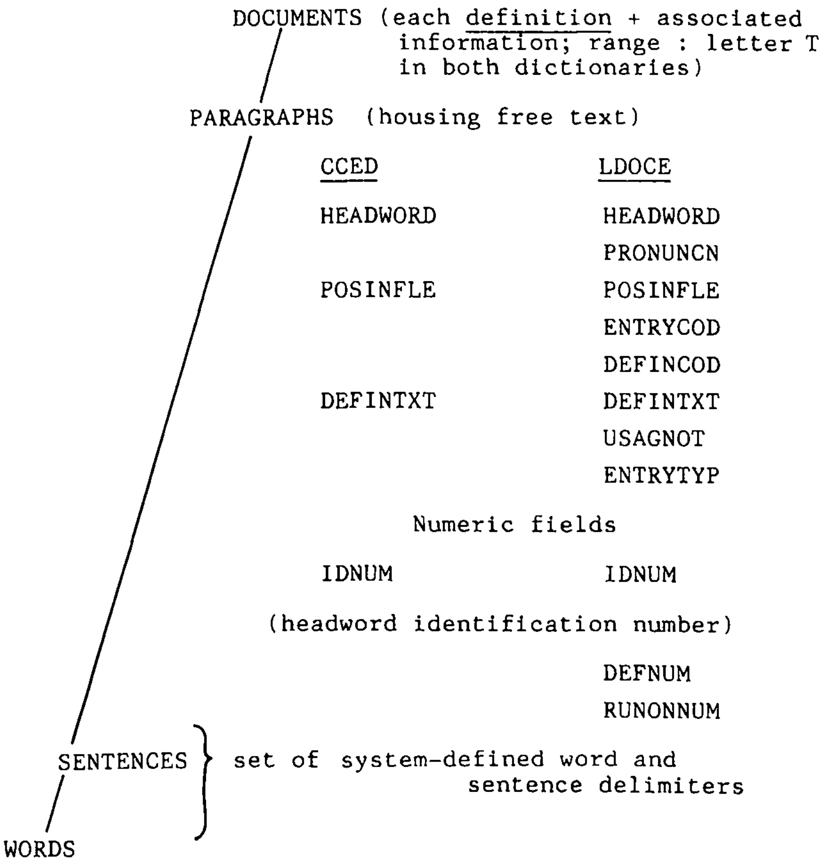
After various clean-up operations performed on the two data-bases, they were ready for input into a STAIRS format. STAIRS is an information-retrieval package produced by IBM, generally used to search text files in a documentation environment. We believe it offers a number of features that make it a useful tool for the investigation and exploitation of dictionary data-bases.

As Fig. 3 shows, STAIRS provides a range of operators for querying a data-base that allow search requests to be formulated as strings. For organizing a corpus into a data-base it provides a hierarchy of units of information ranging from words to documents, that is, from typographically defined linguistic units (words and sentences) to information structures (documents and paragraphs).

Standard results of STAIRS searches are screen displays or print-outs of the relevant documents and statistics on the number of occurrences of each word in the search request and of the number of documents in which it appears. The less familiar by-product of data-base creation under STAIRS is a dictionary and word count of all the word-forms in all the documents. This dictionary can be accessed by merging a given data-base with the system-generated NULL data-base and then requesting the 'new' words, i.e. all words since the NULL data-base is by definition empty.

Fig. 3 Applying the STAIRS package to the LDOCE and CCED data-bases

STAIRS hierarchy



DOCUMENTS (each definition + associated information; range : letter T in both dictionaries)

PARAGRAPHS (housing free text)

CCED	LDOCE
HEADWORD	HEADWORD
	PRONUNCN
POSINFLE	POSINFLE
	ENTRYCOD
	DEFINCOD
DEFINTXT	DEFINTXT
	USAGNOT
	ENTRYTYP

Numeric fields

IDNUM	IDNUM

(headword identification number)

	DEFNUM
	RUNONNUM

SENTENCES

WORDS

set of system-defined word and sentence delimiters

Such STAIRS-produced dictionaries are useful aids for checking to what extent the constraints of a controlled defining vocabulary have been adhered to, for assessing the coverage of such a vocabulary in a dictionary like LDOCE, and for documenting the lexicographers's practice in the use of controlled vs. free defining vocabularies.

Fig. 4 CCED and LDOCE defining vocabularies: extract from Letter J

LDOCE STAIRS dictionary	LDOCE defining vocabulary	CCED STAIRS dictionary
jewellery (3)	jewellery (3)	jewellery
jewels (3)	-	-
jig (1)	-	-
-	-	jilt
-	-	jingle
jitters (1)	-	-
job (31)	job (36)	job
-	-	jobber's
jobs (5)	-	-
-	-	jocular
join (2)	join (6)	join
joined (3)	-	joined
joins (1)	-	joins
joint (4)	joint (5)	joint
-	-	jointed
jointless (1)	-	-
-	-	jointly
-	-	joist
-	-	joists
joke (2)	joke (5)	joke
jokes (2)	-	-
joking (1)	-	-
-	-	jostle
jot (1)	-	jot
-	-	journal
-	-	journalist
journey (21)	journey (24)	journey

Fig. 4 tabulates three dictionaries for words beginning with the letter J:

(1) internal STAIRS dictionary for LDOCE sample - the figures in brackets give the number of occurrences in the data-base;

(2) LDOCE controlled defining vocabulary (cf. LDOCE, p.1285) - the figures in brackets give the number of occurrences in the LDOCE data-base accounted for by the item and its inflexional variants;

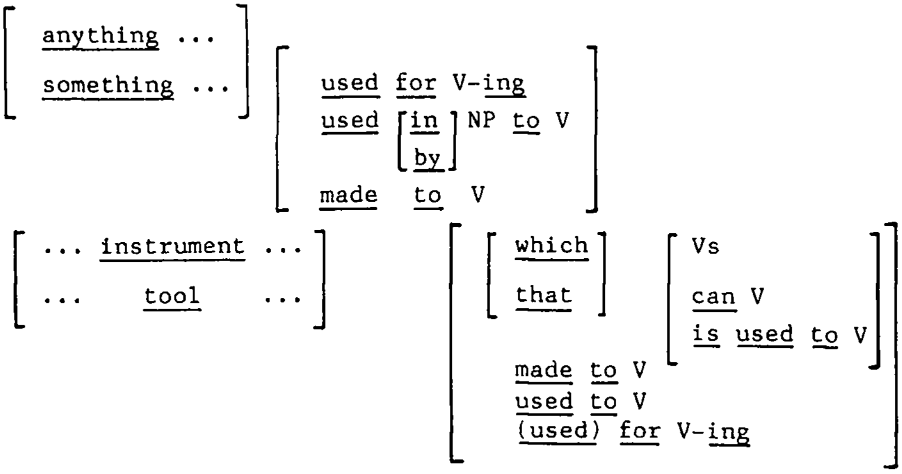
(3) internal STAIRS dictionary for CCED sample.

In the application presented here, we created two STAIRS data-bases for the letter T sample files in LDOCE and CCED. It would have been too costly to process the whole of these two dictionary files and our choice of letter T was largely arbitrary (we had a hunch that letter T contained a sizeable number of 'instruments').

A more crucial decision concerned the entity that we wished to associate with the concept of a STAIRS document. Since a great deal of the information in LDOCE and CCED is word-sense-bound rather than entry-bound, it was decided to regard each LDOCE or CCED definition as a document. In order not to lose the information pertaining to the entry as a whole, J. Jansen wrote a PL/1 program to carry over such information from the entry to each of its definitions.

We can now turn to how we designed the STAIRS queries and to a brief presentation of some of our observations on the results of these queries. As shown in Fig. 5, Michiels (1982) identified a number of patterns which were posited to be highly productive in expressing the 'instrument'-'process' link in LDOCE definitions.

Fig. 5 Action-instrument defining patterns in LDOCE (from Michiels 1982:188)



anything ...
something ...

used for V-ing
used in / by NP to V
made to V

... instrument ...
... tool ...

which / that Vs / can V / is used to V
made to V
used to V
(used) for V-ing

The attempt to turn such patterns into STAIRS queries reveals a number of problems:

(1) STAIRS does not offer facilities for right-to-left processing; in particular it does not allow masking of the left-hand part of a word. For example, the V in a V-ing expression cannot be masked.

(2) Consequently, if we wish to capture the fact that used is optional in the pattern used for V-ing, we are left with for as a single-word query. Words like for and to are heavy-load function words which it is unrevealing to use as single-word queries (see Fig. 6).

Fig. 6 Number of occurrences of for and to in letter T in LDOCE and CCED (from STAIRS dictionaries)

	LDOCE	CCED
for	812	805
to	3909	2674

(3) The least retrievable patterns of all are those which express functional information by means of a relative clause with a finite verb or a verb preceded by can or a near-synonym. If the genus word does not belong to the thesaurus set instrument, tool, ... but is a pro-form (anything, something ...) or some other noun (as in the entries E_1 and E_3 from CCED and E_2 and E_4 from LDOCE) the pattern does not provide any anchor-point for a STAIRS search.

E_1: tabulator ...(2) Computers. a machine that reads from one medium, such as punched cards, producing lists, tabulations, or totals.

E_2: trigger ...(1) the small tongue of metal pressed by the finger to fire a gun

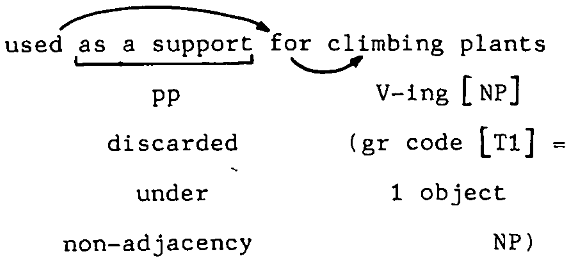
E_3: tap ...(1) a valve by which a fluid flow from a pipe can be controlled...

E_4: tap ... (1) any apparatus for controlling the flow of liquid or gas from a pipe, barrel, etc.

(4) At least one pattern should be added to the set in Michiels (1982), namely used as NP: compare the two readings below for the definition of trellis in entry E_5 from LDOCE.

E_5: trellis ... a light upright framework of long narrow pieces of wood, esp. used as a support for climbing plants ...

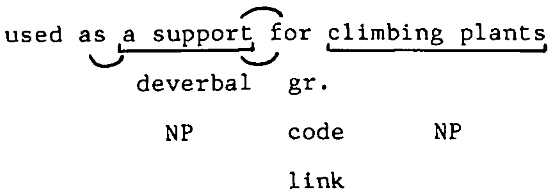
(a) reading under non-adjacency of used for V-ing pattern:


used as a support for climbing plants

pp discarded under non-adjacency

V-ing [NP] (gr code [T1] = 1 object NP)

(b) reading with recognition of used as NP as a defining pattern for instrument-action link:


used as a support for climbing plants

deverbal NP

gr. code link

NP

Extending the STAIRS searches to a data-base that does not feature a controlled defining vocabulary poses the problem of capturing the relevant thesaurus sets. Whereas the 2000-word defining vocabulary in LDOCE can be exhaustively searched for the items belonging to the 'instrument' set (viz. apparatus, instrument, machine, machinery, means, system, tool), such a set is of course open in the case of a free defining vocabulary as that used in CCED. The only procedure is to draw up a list of thesaurically related words (appliance, device, implement, utensil, ...) and check for their occurrence and frequency in the STAIRS-produced dictionary of the data-base (see Fig. 7).

Fig. 7 suggests two striking observations:

(1) heavy use of apparatus in LDOCE and of instrument in CCED;
(2) basically, LDOCE works with a four-term system (apparatus, instrument, machine, tool) and CCED with a five-term system (the above + device).

Predictably, the heavy use in LDOCE of apparatus as genus word has led to some awkward definitions, as in the entries tap (E_4 above) and tin opener (apparatus for opening tins).[2]

We can now move on to the last part of this paper and present some preliminary observations on the use made in CCED and LDOCE of the patterns displayed in Fig. 8.

(1) The first thing to observe is that, irrespective of the policy whether or not to use a controlled defining vocabulary, patterns such as used for, used to, used in, used as are highly productive in both our data-bases. This is probably due to the conjunction of two factors,

Fig. 7 Thesaurus sets in LDOCE AND CCED

	LODCE	CCED
apparatus	31	5
appliance	-	-
device	-	65
equipment	-	9
implement	-	3 (one of which in CCED def. of tool)
instrument	30	59
machine	31	31
machinery	5	4
mechanism	-	8
means	113 (by means of!)	34
system	20	72
tool	15	19
utensil	-	1 (utensils)
designed	-	15
used	489	373

Fig. 8 STAIRS searches on LDOCE and CCED

Single items

apparatus
device
instrument
machine
machinery
means
system
tool

Patterns

made adj. for
made with for

made adj. to
made with to

used adj. as
used with as

used adj. for
used with for

used adj. in
used with in

used adj. to
used with to

(a) the lexicographical tradition which favours a limited number of defining formulae for key semantic relationships and (b) the salience of functional information, as discussed in Labov (1973), Miller (1977), and Clark and Clark (1979) and as evidenced by the fact that it is associated with a very broad spectrum of words, ranging from words denoting typical instruments (*tablespoon*, etc.) to words characterized by such *genus terms* as 'liquid' (e.g. *tabasco*) and 'substance' (e.g. *toothpaste*).

(2) The use of a restricted defining vocabulary does not seem to affect as much as expected the choice of genus term. Here too the weight of the lexicographical tradition is clearly to be felt (cf. Rey-Debove 1971): preference is given in both LDOCE and CCED to fairly general genus terms such as *substance* or *part*; however, the choice of genus word can *affect* the *place* at which functional information is provided in the dictionary. Compare the definitions in the LDOCE and CCED entries E_6 and E_7.

E_6: *tabasco* ... a very hot-tasting liquid ... used *for giving* a special taste to food

E_7: *tabasco* ... a very hot red sauce made from *matured* capsicums

(3) Ideally, the choice of a given pattern in preference to the others (e.g. *used for* V-*ing* rather than *made to* V) should be clearly *motivated*. In many cases, however, *we* at least are unable to see why the lexicographer has opted for a given construction. In principle, for instance, the choice between *used for* V-*ing* and *used in* V-*ing* should be fairly clear: *used for* should point *to* a *more* direct link between 'process' *and* 'instrument' than that indicated by *used in*.

Accordingly we find the following entries (E_8 and E_{10} from LDOCE, E_9 and E_{11} from CCED):

E_8: *tablespoon* ... a large spoon used *for* serving *food ...*

E_9: *tablespoon* ... a spoon ... used *for* serving *food, etc.*

E_{10}: *thread* ... very fine cord ... used *in* sewing or weaving

E_{11}: *thread* ... a fine cord ... used *in* sewing, *etc.*

But we also find E_{12} in LDOCE:

E_{12}: *tuning fork* ... a small steel instrument ... *used* *in* tuning musical instruments.

In addition, one should note the potential ambiguity of *in V-ing*, where the *ing*-form can be either fully verbal or partly or fully nominal (fields such as *gardening* and *weaving*). Lastly, *in V-ing* alternates with *in* + deverbal noun, and it is not clear that the choice of a deverbal, available to CCED but not to LDOCE on account of its restricted defining vocabulary, is motivated.

Notes

[1] We gratefully acknowledge the help of Jacques Jansen, software engineer at the Liège University Computer Centre.

[2] Incidentally, both CCED and LDOCE have the rather unhappy choice of *instrument* as genus word for *tongs* (rightly classified as a *tool* in the OXFORD ADVANCED LEARNER'S DICTIONARY and, significantly, *implement* - a lexicographer's word - in the SHORTER OXFORD ENGLISH DICTIONARY).

References

Clark, E.V. and Clark, H.H. (1979) "When nouns surface as verbs" *Language* 55, 4: 767-811

Labov, W. (1973) "The boundaries of words and their meanings" in *New Ways of Analyzing Variation in English* ed. by C. Bailey and R. Shuy. Washington, D.C.: Georgetown U.P.

Michiels, A. (1982) *Exploiting a Large Dictionary Data Base*. Ph.D. dissertation, Université de Liège

Miller, G.A. (1977) "Practical and lexical knowledge" in *Thinking* ed. by F. Johnson-Laird and P. Wason. Cambridge: U.P.

Rey-Debove, J. (1971) *Etude linguistique et sémiotique des dictionnaires français contemporains*. The Hague: Mouton