WEBSTER'S SEVENTH NEW COLLEGIATE DICTIONARY: A COMPUTER-READABLE FILE FORMAT

James L. Peterson

Department of Computer Sciences
University of Texas at Austin
Austin, TX 78712

TR-196 May 1982

ABSTRACT

A machine-readable version of Webster's Seventh New Collegiate Dictionary exists. This report describes the content and format of the dictionary files.

1.0 INTRODUCTION

The Department of Computer Sciences has acquired a copy of Webster's Seventh New Collegiate Dictionary [1] in a computer-readable form. This is not just a word list, but a copy of the entire dictionary including definitions, cross references, variants, synonyms and so on. It consists of some 12,242,868 characters, with 68,766 main entries. It can be used for all forms of text processing, including spelling, hyphenation, syntax, semantics, and so on. It is available on the departmental PDP-11/60, under UNIX, and also could be written onto magnetic tape for transport to other systems.

The original dictionary was keyboarded onto the Q-32 computer at System Development Corporation (SDC) for a project headed by John Olney [2]. The dictionary was then heavily edited and moved onto an IBM 360. Magnetic tapes of this form were moved to IBM T. J. Watson Research Center and further processed by C. Alberga. A copy of this was acquired by Robert Amsler, then a graduate student at the University of Texas, who used the Pocket Dictionary for his dissertation [3]. We have acquired a copy of the collegiate dictionary from Amsler, and have modified it in many minor ways. This document describes our version.

The dictionary is such a large collection of text that it is broken into 220 files for easier handling. These files reside in the /dictionary directory under the names d.101, d.102, d.103, ..., d.320. The file names were selected to require three digits in all cases.

An index file, d.index, lists the first and last words of each dictionary file. The first letter of all words in a given file is the same; that is when we switch from words starting with "D" to words starting with "E", we start a new file. Otherwise, files are broken to create roughly equal-sized files (from 50,000 to 60,000 characters).

Figure 1 shows a sample of one of the dictionary files. Each line of the file has a character in column 1 which identifies the type and format of the line. The following table shows the number and meaning of each line type.

Line type	Number	Meaning
F	68,766	First line, start for a new word
V	9,957	Variant
D	140,500	Definition, one per line
R	19,123	Related word
X	4,598	Cross-Reference
L	11,990	Label
S	834	Synonym block

Each line is composed of a number of fields. Fields are separated by a semicolon and are defined by their position. The first field of each line is the line type character (F, V, D, L, R, X, or S, as given above). The remaining fields depend upon the type of the line. For example, the second entry (F2) on an F-line is a main entry word; the fifth field (F5) has hyphenation information and the seventh has part-of-speech information.

F;chase;1;;;vb;;

D;1;a;;vt;to follow rapidly : [mini PURSUE]

D;1;b;;vt;[mini HUNT]

D;1;c;;vt;to follow regularly or persistently#
with the intention of attracting or alluring

L;2;;;[italic obs]

D;2;;;vt;[mini HARASS]

D;3;;;vt;to seek out

D;4;a;;vt;to cause to depart or flee : [mini DRIVE]

L;4;b;;[italic slang]

D;4;b;;vt;to take (oneself) off

D;1;;;vi;to chase an animal, person, or thing

D;2;;;vi;[mini RUSH], [mini HASTEN]

S;[bold syn] [mini PURSUE], [mini FOLLOW], [mini TRAIL]:#
[mini CHASE] implies going swiftly after and trying to#
overtake something fleeing or running; [mini PURSUE] suggests#
a continuing effort to overtake, reach, attain; [mini FOLLOW]#
puts less emphasis upon speed or intent to overtake and may not#
imply an awareness on the part of the leader that he is#
pursued; [mini TRAIL] may stress a following of tracks or#
traces rather than a visible object

F;chase;2;;;n;;

D;1;a;;n;the act of chasing : [mini PURSUIT]

D;1;b;;n;[mini HUNTING] -- used with [italic the]

D;1;c;;n;an earnest or frenzied seeking after something desired

Figure 1. Sample of the dictionary file.

The table below shows the contents of the fields for each line type. Appendix A discusses each line and field in more detail.

F Card

F1: F
F2: Main Entry
F3: Homograph Number
F4: Prefix/Suffix/Infix
F5: Hyphenation
F6: Part of Speech
F7: Part of Speech Joiner
F8: Secondary Part of Speech

V Card

V1: V
V2: Variant Word
V3: Hyphenation
V4: Variant Level

D Card

D1: D
D2: Sense Number
D3: Sense Letter
D4: Sense Subnumber
D5: Part of Speech
D6: Definition Text

R Card

R1: R
R2: Related Word
R3: Hyphenation
R4: Part of Speech
R5: Part of Speech Joiner
R6: Secondary Part of Speech

X Card

X1: X
X2: Word
X3: Superscript
X4: Subscript
X5: Type of Cross Reference
X6: Secondary Word

L Card

L1: L
L2: Sense Number
L3: Sense Letter
L4: Sense Subnumber
L5: Label Text

S Card

S1: S
S2: Synonym Text

Some lines, particularly definitions and synonym blocks are quite long and hence it is difficult to fit them on one 80 character line. Therefore, lines are split whenever necessary so that no physical line exceeds 80 characters. Lines are always split at a blank, and the incomplete line is terminated with a sharp or hash mark character ("#"). When processing the dictionary, if a line is terminated with a # character, then the # character should be replaced by a blank and the next physical line should be read in and appended to the previous line.

An examination of the order of occurrences of the various types of lines reveals that the following are the most common forms of dictionary entries: FD, FDLD, FDLDLD, FDLDR, FDR, FDRD, FDRV, FDRVR, FDS, FDX. Each entry starts with an F-card, and is normally followed by a sequence of D-cards. The longest entry has 215 cards.

2.0 CHARACTER CODES

A major problem with the dictionary is its character set. First, the dictionary publisher did not feel constrained in his use of characters, but choose whatever symbols best fit his purpose. Second, the dictionary was originally encoded in an extended BCD (for the Q-32 computer), then translated into EBCDIC (for the IBM 360/370) and now has been translated into ASCII (for our PDP-11/60). None of these character sets is completely compatible with the others, nor is any of them sufficient to represent all of the variation found in the original printed dictionary. Hence an encoding scheme must be used to expand the set of representable characters. This expansion occurs in two independent directions: font information and special characters.

2.1 Font Information

Font information is represented by use of the square brackets in ASCII to surround any special font material. Five font types are recognized: (1) italic, (2) mini-caps, (3) bold, (4) subscripts, and (5) superscripts. Each is denoted by an identifying keyword immediately after the opening left square bracket, followed by a space, followed by the material to be in the defined font, followed by the closing right square bracket.

italic	[italic ...]
bold	[bold ...]
mini-caps	[mini ...]
subscripts	[sub ...]
superscripts	[sup ...]

For example, an italic "was" is represented as "[italic was]", while a mini-caps "AMBIENT" is "[mini AMBIENT]" and a bold "syn" is "[bold syn]". Superscripts and subscripts may be italic, mini-caps, or bold, and a few superscripted superscripts also occur, as in 6.24 {times} 10 [sup 10 [sup 10]].

2.2 Special Symbols

The dictionary includes a large number of special symbols which are not representable in ASCII. These include all the Greek alphabet, the Hebrew alphabet and many miscellaneous other special symbols. All special symbols which are not available in ASCII (and some that are) have been given names which are represented by the name encased in braces, as {degrees} (for a degree symbol), {times} (for multiplication represented by a small x), {tau} (for the lower-case greek letter tau), and so on. A complete list of these special symbols is in Appendix B.

Each symbol name has been selected to exclude embedded blanks. Thus all characters between an opening right brace and its closing right brace are non-blank. Certain characters which are in ASCII (braces, brackets, question mark, exclamation mark, and so on) have also been represented in this way because it allowed them to be used for other purposes (such as font and special character representation), and they occurred only infrequently (less than 100 times).

3.0 ERRORS IN W7

While processing W7 to both understand its contents and put those contents into a useable form, we encountered a large number of errors. These errors were of several types:

1. Merged Illustrations. For example, under "false", the illustration was "< ~ documents ~ teeth >" and should have been "< ~ documents> < ~ teeth>". To correct this we searched for any line of the form "< ... ~ ... ~ ... >".

2. Accents (236 entries with accents). The accent field was wrong about half the time. The normal problem was that the accent was on wrong letter. In these cases, the hyphenation information generally showed syllables that were 2 letters too long.

3. Incorrect values in fields. We created a sorted list by frequency of the contents of each field (as listed in Appendix A). These could then be examined for rare or inappropriate values; for example a ´g´ in a numeric field, or a zero in an alphabetic field.

4. Mismatched parentheses or brackets. We wrote a program to simply count parenthesis, braces and brackets. Many were found to be mismatched.

All of these errors, once found and verified, were corrected by hand, using the text editor.

Another form of error analysis was an attempt to detect typographical errors. The approach was quite simple: we extracted a list of all unique words used in the dictionary definitions. This produces a list of 54,298 words. We compared this list with the list of

all words defined in the dictionary (main entries, variants or related words). This reduced our list to 20,292 words which were used in definitions, but not defined. Many of these were derived forms of defined words: past tense, plurals, and so on. Doing some simple suffix analysis, we were left with about 8,000 words. Most of these were apparently Greek or Latin botanical, zoological or geographical names. Deleting those ending in "-ia" or "-ae" and all words which were in italics in the dictionary left a list of 2821 words.

These words were checked by hand to produce a list of 903 incorrectly spelled words. We also found 54 words which were used, but not defined.

Australasian	hindlimbs	rosebush
bloomery	homeward	sawteeth
broadheaded	hyperactivity	seneschal
clothesline	leftward	sheepdogs
crossbeam	lightcolored	shorthaired
darkskinned	longlegged	sightsee
dinnerware	lowcut	snowstorm
doorbell	messroom	songbook
entranceway	Mr.	spinymargined
equivalve	nailhead	spondumene
etc	neckband	sulphates
fieldwords	noctuid	supersensitized
flattish	nubby	tv
foreseen	parimutuel	twelvefold
foretold	partaken	understock
fourpence	pregenital	upcurved
gunstock	pyrotechny	valency
hairdress	rangeland	workpiece

We also found a smaller list of words with typographical errors in the main entry in the computer files.

Of the 903 typographical errors, 543 were the result of a missing blank between two words. Of the remaining 360 typographical errors:

34% were a missing letter
27% were a wrong letter
20% were an extra letter
13% were transposed letters

The remaining errors were caused by 2 extra or 2 missing letters, or by transposing two letters around a third.

We also found ten cases of typographical errors in the original printed dictionary.

1. "apostasize" should be "apostatize" in "lapse".

2. "clubmosses" should be "club mosses" in "lycopodium"

3. "doublecross" should be "double-cross" in "bitch"

4. "goodhumored" should be "good-humored" in "kid"

5. "knicknack" should be "knickknack" in "vanity"

6. "permanenty" should be "permanently" in "type species"

7. "quantitive" should be "quantitative" in "drift"

8. "genu" should be "genus" in "capsicum"

9. "gulley" should be "gully" in "barranca"

10. "NCE" should be "New Catholic Edition" in "fornication"

A further investigation has been made into the correctness of the hyphenation information provided. By comparing the defined hyphenation points with the words, we found several entries where the hyphenation points were clearly incorrect. For example "pi" was defined to have a hyphenation point after the third character. There were also a list of 13 words found which were misinterpreted. These were variant or related words. The dictionary does not give hyphenation for an entry more than once, even if the word has several main entries. However, for variant and related words, it may be necessary to hyphenate the word as an artifact of the narrow column width of the dictionary. In several cases, this hyphenation was misinterpreted as a hyphenation definition, rather than a hyphenation use.

For example, consider the hyphenations of "futilitarian". As a main entry it is listed as "fu-til-i-tar-i-an", but as a related adjective, it is listed as "fu-tilitarian". This second hyphenation is a use of hyphenation, not a definition.

Additional error checking is ongoing. Our experience has indicated that with a document this large, it is probable that there will always be errors. These errors are being corrected as they are encountered.

4.0 USE OF THE DICTIONARY

The existence of the W7 dictionary makes possible a large number of text and word analysis functions. However, the size of W7, plus its representation, make the use of the information which is available very difficult. We have written over 40 programs to manipulate the W7 dictionary. These programs allow us to extract fields, search for patterns, modify the representation format, and generally find and transform information into more useful forms. This report has been prepared, in part, to aid in the production of further programs by indicating the range of values for the various fields in the dictionary.

APPENDIX A

Format of the Dictionary Files

1.0 F LINES - MAIN ENTRY

1.1 Field F1 - F

1.2 Field F2 - Main Entry

This field contains a main entry: a word or phrase, which is defined in the dictionary. Generally, this is a lower-case word, but entries include,

1. Proper names and trademarks with an initial upper case letter (e.g., "Paleocene", "George", "Angledozer").

2. Phrases with embedded blanks and possibly embedded upper case letters (e.g., "gray matter", "run away", "Binet-Simon scale", "Magellanic Cloud").

3. Compound words with embedded hyphens (e.g., "black-and-blue", "battle-ax", "bird's-foot trefoil").

4. Prefix and suffix entries with trailing and leading hyphens (e.g., "bath-", "bi-", "-biosis", "-ship").

5. Contractions and possessives with embedded, leading, or trailing apostrophes (e.g., "'m", "mayn't", "rock 'n' roll", "Adam's", "Mohs' scale").

6. Numeric entries and mixed alphanumeric entries (e.g., "A1", "M-1", "carbon 14", "1080" (ten-eighty)).

7. Embedded special characters (such as commas and periods) (e.g., "Mrs. Grundy", "Court of St. James", "lock, stock, and barrel", "Tom, Dick, and Harry").

8. Non-ASCII characters (e.g., "ch{a^}teau", "ca{n~}on", "k{u:}mmel").

9. Typesetting commands (e.g., "vitamin B[sub 12]", "F[sub 1] layer").

1.3 Field F3 - Homograph Number

This field is normally empty, but may contain a number (from 1 to 9) representing a homograph number. The following list indicates the number of entries with each homograph number. An entry with a maximum homograph of 5 will have entries with homograph numbers of 1, 2, 3, 4, and 5.

homograph number	number of entries
1	6538
2	6542
3	1427
4	475
5	164
6	54
7	17
8	5
9	3

The two entries with exactly 8 homographs are "bob" and "like"; the three entries with 9 homographs are "list", "rack" and "tip".

1.4 Field F4 - Prefix/Suffix/Infix

For prefix and suffix entries, this field has either a "p" (for prefix), "s" (for suffix) or "i" (for infix). Normally, it is empty.

field character	number of entries
p	663
s	444
i	2

The two entries with an "i" value are "-i-" and "-o-".

1.5 Field F5 - Hyphenation

This field contains hyphenation information. This is a sequence of one-digit numbers giving the number of characters between possible hyphenation points. As an example, a word such as "estimate" is hyphenated as "es-ti-mate", and would be encoded as "22". The distance from the start of the word to the first hyphenation point is two, and so is the next syllable. The word "ethnological", hyphenated as "eth-no-log-i-cal", is encoded as 3231.

The encoding of the distance between hyphenation points is one digit, from 1 to 9. If the distance exceeds 9, we continue with the upper case letters (as with hexadecimal). Thus A is 10, B is 11, C is 12, ..., Z is 35.

There are 2,550 distinct values for this field. The most common entries are:

hyphenation encoding	number of entries
3	6707
4	4723
2	3600
23	2452
32	2372
33	2242
22	1911
5	1767
24	1086

The frequency of occurrence for each of the individual characters is:

Span	Number of Occurrences	
1	6722	
2	30603	
3	30044	
4	11455	
5	3429	
6	604	
7	375	
8	307	
9	180	
A	38	
B	49	
C	28	
D	12	
E	4	
F	5	
G	1	"chuck-will's-widow;G"
I	1	Tom, Dick, and Harry;I

The longest hyphenation encoding is

pneumonoultramicroscopicsilicovolcanoconiosis;422232342312322221

pneu-mo-no-ul-tra-mi-cro-scop-ic-sil-i-co-vol-ca-no-co-ni-o-sis

1.6 Field F6 - Part Of Speech

This field identifies the primary part of speech of the main entry. The part of speech is encoded in a two-character mnemonic abbreviation. The parts of speech and their frequency are listed below.

POS	Frequency	Meaning and Examples
n	42587	noun
aj	13253	adjective
vt	4974	verb, transitive
vb	2422	verb
av	1432	adverb
vi	1362	verb, intransitive
cf	516	combining form
pp	164	preposition
nc	150	noun combining form
tm	121	trademark
pn	108	pronoun
ns	107	noun suffix
cj	96	conjunction
ij	94	interjection
pf	74	prefix
js	50	adjective suffix
vs	12	verb suffix
vp	10	verb imperative
as	7	adjective suffix
va	4	verbal auxiliary
ia	3	indefinite article
vm	2	verb impersonal, past "meseems", "methinks"
vc	2	verb combining form "-lyze", "-sect"
sf	2	suffix "-est", " -fold"
da	2	definite article "the", "ye"
np	1	noun plural suffix "-s"
is	1	interjection suffix "-o"

1.7 Field F7 - Part Of Speech Joiner

For those cases where two parts of speech are given for a single main entry (most separate parts of speech rate separate main entry listings), this field indicates how the part of speech given in field F9 is related to the part of speech of field F7.

Relationship	Encoding	Number
Equal	1	48
Secondary	2	205

One anomoly is the entry "vires" which has an F7 and F8 field, but an empty F6 field.

1.8 Field F8 - Secondary Part Of Speech

When a secondary part of speech is given, it is listed in field F8. The relationship between this part of speech and the part of speech listed in field F6 is either equal or secondary, and is indicated by the contents of field F7.

Part of Speech	Frequency		
aj	182	adjective	
av	34	adverb	
n	23	noun	
vb	3	verb	
vt	2	verb, transitive	
js	2	adjective suffix	
cj	2	conjunction	
as	2	adverb suffix	
vi	1	verb, intransitive	"bike"
pp	1	preposition	"betwixt"
pn	1	pronoun	"-'s"

If you look at the number and combination of F6 and F8 fields, we find the following chart of pairs with more than one occurrence:

F6;F8	number
av;aj	171
aj;av	34
aj;n	19
n;aj	7
pn;aj	3
n;vb	3
vs;js	2
n;vt	2
js;as	2
av;cj	2

2.0 V LINES - VARIANTS

2.1 Field V1 - V

2.2 Field V2 - Variant Word

This field contains the variant word or phrase. There are 9957 such entries. As with the F2 field, there are several special cases that may need special attention.

1. Proper names with an initial upper case letter (e.g., "Flatheads").

2. Phrases with embedded blanks and possibly embedded upper case letters (e.g. "Magna Carta", "amici curiae").

3. Compound words with embedded hyphens (e.g., "counter-reformation").

4. Prefix and suffix entries with trailing and leading hyphens (e.g., "astro-", "chemo-", "-bioses", "-metrical").

5. Contractions with embedded, leading, or trailing apostrophes (e.g., "M-1's", "objets d'art", "one's self").

6. Numeric entries and mixed alphanumeric entries (e.g., "606", "400", "21").

7. Non-ASCII characters (e.g., "neglig{e'}", "{o:}re", "20{diagonal}20").

8. Typesetting commands (e.g., Vitamin B[sub 1]).

2.3 Field V3 - Hyphenation

The hyphenation encoding is as described for field F5. The syllable lengths run from 1 to 9, and also include a few entries of A, B, C, D, and F. The F value is for "jack-in-the-boxes;F". The longest entries ("onomatopoetically;2142213" and "thermoperiodicity;4222131") are 7 digits in length.

2.4 Field V4 - Variant Level

This field encodes the relationship of the variant to its main entry. The encoding is a two-digit code. The first digit indicates the entry of which this is a variant. A "0" indicates that this is a variant of a main entry; a "1" indicates that this is a variant of the first inflected form (such as the comparative form of an adjective), and so on. The second digit indicates (with a "1") if the two forms are equally acceptable alternates, or (with a "2") if this variant form is a secondary variant.

Encoding	Frequency	
11	5093	irregular plural (e.g. "ganglia")
01	2460	variant spelling (e.g. "intortion")
21	1227	present participle (e.g. "chagrining")
02	661	secondary spelling (e.g. "bivalved")
12	322	(e.g. "birle" under "birl")
31	99	(e.g. "were" under "be")
22	69	(e.g. "bayoneting" under "bayonet")
32	11	(e.g. "awaking", "cleaving", "thriving")
41	6	(e.g. "flied", "goes", "up")

B1	1	("be")
A1	1	("are")
91	1	("is")
81	1	("are")
71	1	("am")
61	1	("being")
51	1	("been")

It is not clear that this field is consistently or correctly defined.

3.0 D LINES - DEFINITIONS

3.1 Field D1 - D

3.2 Field D2 - Sense Number

Many words have many different meanings, or senses. Each different sense if given a sense number: the first sense is sense 1, and so on. Where an entry has only one sense, it is either sense zero or sense one.

Sense Number	Frequency	
0	46191	
1	36413	
2	33867	
3	11956	
4	5204	
5	2682	
6	1559	
7	872	
8	532	
9	335	
10	233	
11	183	
12	143	
13	97	
14	66	
15	47	
16	29	
17	22	
18	19	
19	13	
21	9	
23	5	
22	5	
20	5	
24	4	
25	1	"set"
26	1	"set"

3.3 Field D3 - Sense Letter

In addition to having multiple senses, some words have significant variations in meaning within a given sense. To distinguish these subsenses, they are given a sense letter in addition to a sense number. Thus, we can refer to sense "3a" as well as sense "3b" and "3c", where appropriate.

Sense Letter	Frequency	
a	16659	
b	16578	
c	4307	
d	1214	
e	409	
f	153	
g	65	
h	30	
i	12	
j	10	
k	4	
l	4	
m	1	"call"
n	2	"call"

3.4 Field D4 - Sense Subnumber

Finally, in addition to having senses and subsenses, there are cases where we have sub-sub-senses. These are indicated by a subsense number. To completely distinguish these cases, one must give a sense number, a sense letter and the subsense number (as in definition "3b1").

Sense Subnumber	Frequency	
1	1736	
2	1739	
3	277	
4	69	
5	19	
6	3	
7	2	"good", "stone"
8	1	"stone"

3.5 Field D5 - Part Of Speech

Although the main entry card (F-card) indicated for each word its part of speech, we actually could have had two parts of speech on the F-card: the part of speech and the secondary part of speech. The D5 field specifies the part of speech for this particular definition. This

will normally be one of the two parts of speech from the F-card. However, if the main part of speech on the F card was a verb ("vb"), then this field may further classify that as either a transitive or intransitive verb.

Part of Speech	Frequency	
aj	26088	adjective
vt	19415	verb, transitive
n	14485	noun
vi	8869	verb, intransitive
av	2320	adverb
cf	863	combining form
pp	552	preposition
vb	372	verb
pn	225	pronoun
pf	225	prefix
ns	223	noun suffix
cj	200	conjunction
nc	188	noun combining form
tm	128	trademark
ij	98	interjection
js	89	adjective suffix
va	65	verbal auxiliary
vs	26	verb suffix
da	21	definite article
vp	11	verb imperative
as	8	adjective suffix
ia	7	indefinite article
sf	4	suffix
vm	2	verb impersonal, past
vc	2	verb combining form
np	2	noun plural suffix
is	1	interjection suffix

3.6 Field D6 - Definition Text

The remainder of the D-card is the definition of the indicated sense of the dictionary entry. The definitions were transcribed directly from the printed text.

4.0 R LINES - RELATED WORDS

4.1 Field R1 - R

4.2 Field R2 - Related Word Or Phrase

The related word entry is similar to the F2 and V2 fields. We have the same sort of special cases.

1. Proper names with an initial upper case letter (e.g. "Fabian").

2. Phrases with embedded blanks and possibly embedded upper case letters (e.g. "glom on to", "High Churchman").

3. Compound words with embedded hyphens (e.g. "globe-trotting", "Indo-Aryan").

4. Suffix entry with a leading hyphen ("-carpy").

5. Contractions with embedded apostrophes (e.g., "in one´s cups", "from the horse´s mouth").

6. Non-ASCII characters (e.g. "clich{e´}", "Zu{n~}ian").

4.3 Field R3 - Hyphenation

Hyphenation information is encoded as with field F5 and V3. The hyphenation spans run from 1 to 9, with a few having values of A, B, C, D, and G. The G value is for "will-o´-the-wispish;G". The longest field for R3 is for "interdenominationalism;23231242".

4.4 Field R4 - Part Of Speech

This field gives the part of speech of the indicated related word, similar to field F6.

Part of Speech	Frequency	
n	7981	noun
aj	5258	adjective
av	4838	adverb
vt	185	verb, transitive
vb	102	verb
vi	96	verb, intransitive
nc	2	noun combining form
ac	2	adjective combining form ("-cratic")
pp	1	preposition ("midst")

4.5 Field R5 - Part Of Speech Joiner

This field is similar to field F7, and indicates the relationship between fields R4 and R6.

Relationship	Encoding	Number
Equal	1	340
Secondary	2	6

The six fields with a value of 2 are related to "AWOL", "crescendo", "darn", "half hour", "pitter-patter", and "slant".

4.6 Field R6 - Secondary Part Of Speech

This field is similar to field F8.

Part of Speech	Frequency	
n	178	noun
aj	112	adjective
pn	34	pronoun
av	20	adverb
vb	1	verb ("cross-reference")
pp	1	preposition ("mid")

5.0 X LINES - CROSS REFERENCE

5.1 Field X1 - X

5.2 Field X2 - Word

The X2 field is similar to the F2, R2, and V2 fields. The entries may contain capital letters, blanks, hyphens, and some special characters (including periods, parenthesis, numbers, and typesetting codes). For each cross reference, the type of cross reference is encoded in field X5. Fields X3 and X4 help to specify exactly to which of the possible multiple senses of the cross-referenced word reference is being made.

5.3 Field X3 - Superscript

If the cross-reference word has homographs, we need to know which is being referred to. This entry matches the F3 value.

Superscript	Frequency	
1	3	"his", "herself", "-s"
3	1	"squabble"

5.4 Field X4 - Subscript

Once a cross-reference entry has been defined, it may be important to specify which of the possibly multiple senses is being referred to. Following the word may be a sense specifier. This can indicate simply a sense number, or a sense number and sense letter. This seems particularly important when a comparison is being made. For example, consider the cross reference to "base" under "alkali" or to "meta-", "orth-" and "para-" under "benzene ring".

Subscript	Frequency
1	5
3b	2
3	2
2b	2
2a	2
1b	2
1a	2
7	1
4b	1
2	1

5.5 Field X5 - Type Of Cross Reference

The original report from SDC [2] indicates that there may be 11 different values for this field; we found that only the following 8 actually occur.

Encoding	Frequency	Meaning	Example
9	2996	[bold syn] see <X2>	"averse"
6	572	called also <X2>	"dipper"
4	393	compare <X2>	"averse"
1	389	see *** table	"aluminum"
0	103	see <X2>	"anth-"
8	90	[bold syn] see in addition <X2>	"abandon"
3	53	see <X6> at [mini MONEY] table	"agora"
5	1	compare [mini ELEMENT] table	"rare earth"

The symbol <X2> means the contents of field X2 on this card; the symbol <X6> means the contents of field X6. The *** of code 1 can stand for "money" or "element" (at least) and is not recorded.

As with some of the other fields, it is not clear that this field has been consistently applied. There are still some occurrences of "called also" in the definitions, although these are generally scattered throughout only those words starting with an "a", "b" or "c". We found 112 occurrences of "called also [italic ...]" in the files, see for example the entries for "Haitian", "hemorrhoid", "hydroxide ion", and "road".

5.6 Field X6 - Secondary Word

Each cross reference with an encoding of 3 in field X5 is of the form "See <X6> at [mini MONEY] table", where <X6> is some form of money. This field contains the form of money. The most common form is "pound" (8 occurrences), followed by two multiple occurrences of "dinar", "yuan", "rupee", "krona", "franc", and "escudo". There is exactly one use of the remainder: "zloty", "yen", "somalo", "schilling", "rupiah", "ruble", "riyal", "riel", "peso", "peseta", "markka", "lira", "lev", "leu", "lek", "kyat", "krone", "koruna", "kip", "hwan", "guarani", "gourde", "forint", "drachma", "deutsche mark", "cruzeiro", "colon", "bolivar", "balboa", "baht", "afghani".

6.0 L LINES - LABELS

6.1 Field L1 - L

6.2 Field L2 - Sense Number

To specify the entry which is being labeled, we must specify the sense number, letter and subnumber as appropriate. In general, it is expected that the definition being labeled would follow the label, but this may not always be the case.

Sense Number	Frequency	
1	2269	
2	1934	
0	1318	
3	658	
4	278	
5	148	
6	99	
7	50	
8	32	
9	20	
10	9	
11	9	
12	9	
13	7	
14	4	
15	3	
19	1	"life"

21	1	"set"
24	1	"strike"

6.3 Field L3 - Sense Letter

This field is similar to field D3.

Sense Letter	Frequency	
b	699	
a	559	
c	211	
d	52	
e	11	
f	4	
h	3	"dead", "matter", "what"
g	1	"matter"

6.4 Field L4 - Sense Subnumber

This field is similar to field D4.

Subsense Number	Frequency	
2	71	
1	49	
3	8	
4	3	"number", "objective", "rate"

6.5 Field L5 - Label Text

The label itself is a short phrase indicating some special property of the entry. Below are the most common properties listed in the labels. These are often combined, as in "chiefly Scot", "often cap M", and "sing but pl in constr".

Keyword	Frequency
pl	3854
archaic	2092
of	1473
often	1239
cap	1127
obs	1115
chiefly	949
Brit	767

attrib	561
var	518
Scot	488
dial	471
sing	403
slang	396
constr	379
past	172
usu	97
Eng	57
part	39
plant	35
Midland	34
pres	33
animal	31
South	30

7.0 S LINES - SYNONYM BLOCK

7.1 Field S1 - S

7.2 Field S2 - Synonym Text

There are 835 S-cards. Each is one line with an average of 480 characters, 72 words. The synonym blocks are of the general form: "S;[bold syn] <list of words>: phrases;". The list of words contains a set of words (in mini-caps) which are similar. The phrases take each word and separately indicate the special meaning of that word, as opposed to the other similar words.

APPENDIX B

Special characters

The following is a list and explanation of the various special characters defined using the brace notation. For symbols which are not in the ASCII character set, or those which occur only infrequently, we have selected a name for the character. The name is enclosed in braces, and should be interpreted as a single character.

The first group of characters are those alphabetic characters which may have diacritical marks, such as a grave accent, acute accent, umlaut, carat, or tilda. These are encoded by placing the alphabetic character and its accent mark in braces.

Below we give the symbol, the number of times that it occurs, and an entry word giving an example. Notice that in some cases (such as the entry for {C,}) the symbol is used not in the example ("compendium"), but in the definition of the example.

9	{a:}	"fr{a:}ulein"
15	{a^}	"p{a^}te"
20	{a`}	"folie {a`} deux"
1	{A`}	"vis-{a`}-vis"
10	{c,}	"cura{c,}ao"
3	{C,}	"compendium"
209	{e´}	"d{e´}class{e´}"
10	{E´}	"work"
6	{e:}	"Chlo{e:}"
26	{e^}	"f{e^}te champ{e^}tre"
2	{E^}	"vis-{a`}-vis"
37	{e`}	"cr{e`}me de la cr{e`}me"
4	{E`}	"calash"
9	{i:}	"na{i:}ve"
3	{I:}	"wide-eyed"
5	{i^}	"ma{i^}tre d´h{o^}tel"
1	{I^}	"ma{i^}tre d´"
19	{n~}	"ma{n~}ana"
11	{o:}	"r{o:}ntgen"
5	{o^}	"table d´h{o^}te"
3	{O^}	"prix fixe"
11	{u:}	"k{u:}mmel"
1	{U:}	"Geiger counter"

Notice that with these characters, both upper and lower case alphabetics are recorded separately.

The larger set of characters are unusual symbols. These occur, in almost all cases, in definitions, often the definition of the symbol itself.

4	{a-e}	"ligature"
2	{accent-acute}	"acute"
3	{accent-grave}	"grave"
1	{aleph}	"aleph"
1	{alpha-with-smooth-breathing}	"smooth breathing"
5	{alpha}	"alpha", "Altair"
4	{asterick}	"asterisk", "ellipsis"
1	{asterism-down}	"asterism"
1	{asterism-up}	"asterism"
1	{ayin}	"ayin"
1	{b-minus-c-overbar}	"vinculum"
1	{b-sup-macron}	"Verner's law"
5	{bar-e}	"high", "lax"
2	{bar-k}	"palatal"
2	{bar-n}	"phonemic"
1	{bar-th}	"fricative"
2	{beta}	"beta"
1	{beth}	"beth"
1	{breve}	"breve"
15	{by}	"cord", "crown", "medium"
1	{c-with-vertical-bar}	"alla breve"
3	{center-dot}	"congruent", "syllable"
1	{check}	"check"
886	{chemical-formula}	"tiemannite", "acetal"
3	{chi}	"chi", "Verner's Law"
1	{CHI}	"chi"
3	{circumflex}	"accent", "caret"
1	{close-double-curly-quotes}	"quotation mark"
1	{close-double-straight-quotes}	"quotation mark"
1	{close-single-curly-quotes}	"quotation mark"
1	{close-single-straight-quotes}	"quotation mark"
1	{curcumflex}	"circumflex"
1	{curly-A}	"swash"
1	{curly-N}	"swash"
1	{curly-P}	"swash"
1	{curly-R}	"swash"
4	{dagger}	"dagger", "Ramism"
1	{daleth}	"daleth"
75	{degree}	"dihedral", "liter"
2	{delta}	"delta"
3	{DELTA}	"delta"
9	{diagonal}	"diagonal", "proportion"
3	{division-symbol}	"obelus"
16	{dollar-mark}	"charge", "go", "tune"
1	{double-arrows}	"equation"
1	{double-bar-dollar-mark}	"dollar mark"
1	{double-dagger}	"double dagger"
1	{double-half-arrows}	"equation"

1	{double-hyphen}	"double hyphen"
2	{edh}	"edh"
1	{epsilon}	"epsilon"
15	{equals}	"equation", "sine curve"
2	{eta}	"eta"
4	{exclamation-point}	"oh", "that", "yes"
1	{fermata-down}	"fermata"
1	{fermata-up}	"fermata"
1	{fist}	"index"
1	{flat-sign}	"flat"
1	{g-sub-macron}	"Verner's law"
2	{gamma}	"gamma"
1	{GAMMA}	"gamma"
1	{gimel}	"gimel"
1	{heth}	"heth"
1	{he}	"he"
1	{horizontal-brace}	"brace"
6	{inches}	"punchboard"
1	{integral-sign}	"sign"
1	{iota}	"iota", "iotacism"
2	{IOTA}	"iota"
1	{kaph(final)}	"kaph"
1	{kaph}	"kaph"
3	{kappa}	"kappa"
1	{lambda}	"lambda"
1	{LAMBDA}	"lambda"
1	{lamed}	"lamed"
1	{left-brace}	"brace"
3	{left-square-bracket}	""fulminate"
17	{linguistic-formula}	"meter", "rough breathing"
1	{macron}	"macron"
1	{mem(final)}	"mem"
1	{mem}	"mem"
8	{minus}	"sign", "vinculum"
18	{minutes}	"ecliptic", "south by east"
1	{mu}	"mu"
1	{natural-sign}	"natural"
1	{nun(final)}	"nun"
1	{nun}	"nun"
1	{nu}	"nu"
1	{o-e}	"diphthong"
1	{OMEGA}	"omega"
1	{omega-with-rough-breathing}	"rough breathing"
3	{omega}	"omega"
1	{omicron}	"omicron"
1	{one-dot-a}	"low"
3	{one-dot-o}	"syllabic"
2	{one-dot-u}	"tense"
1	{open-double-curly-quotes}	"quotation mark"
1	{open-double-straight-quotes}	"quotation mark"
1	{open-single-curly-quotes}	"quotation mark"
1	{open-single-straight-quotes}	"quotation mark"
1	{paragraph-mark}	"paragraph"
1	{parallel}	"parallel"
1	{pe(final)}	"pe"

1 {pe}	"pe"
1 {phi}	"phi"
1 {PHI}	"phi"
6 {pi}	"pi", "number"
1 {PI}	"pi"
2 {plus-or-minus}	"functor", "square root"
11 {plus}	"can", "independent"
1 {presa}	"presa"
1 {prime}	"prime"
1 {psi}	"psi"
1 {PSI}	"psi"
1 {qoph}	"qoph"
11 {question-mark}	"do", "oh", "so"
1 {radical-sign}	"radical sign"
1 {radical}	"sign"
1 {resh}	"resh"
1 {rho-with-rough-breathing}	"rough breathing"
4 {rho}	"rho", "rough breathing"
1 {right-arrow}	"equation"
1 {right-brace}	"brace"
3 {right-square-bracket}	"bracket"
2 {rough-breathing-mark}	"rough breathing"
2 {sadhe(final)}	"sadhe"
2 {sadhe}	"sadhe"
1 {samekh}	"samekh"
5 {schwa}	"schwa", "allomorph"
2 {seconds}	"obliquity"
1 {section-mark}	"section"
4 {sharp}	"sharp", "whole step"
1 {shin}	"shin"
1 {sigma-1}	"sigma"
2 {sigma-2}	"sigma"
1 {SIGMA}	"sigma"
1 {sin}	"sin"
1 {slur-down}	"slur"
1 {slur-up}	"slur"
2 {smooth-breathing-mark}	"smooth breathing"
1 {space-mark}	"space mark"
1 {spade}	"spade"
1 {squareroot-of-3}	"surd"
2 {squareroot-of-minus-one}	"complex number"
3 {tau}	"tau", "tenuis"
1 {taw}	"taw"
1 {ten-to-ten-to-hundred}	"googolplex"
1 {teth}	"teth"
1 {theta}	"theta"
1 {THETA}	"theta"
2 {thorn}	"thorn"
1 {tilda}	"circumflex"
8 {times}	"demy", "fatshedera"
2 {umlaut}	"diaeresis", "umlaut"
1 {upsilon}	"upsilon"
1 {UPSILON}	"upsilon"
1 {waw}	"waw"
2 {xi}	"xi"

1 {XI}	"xi"
1 {yod}	"yod"
1 {yogh}	"yogh"
1 {zayin}	"zayin"
1 {zeta}	"zeta"

There are also a bunch of fractions, which are of the form {a/b} where a and b are integers. All fractions are used infrequently. {1/2} is used 26 times, {1/4} 12 times, {3/4} and {1/3} 9 times, {1/20} 7 times.

{1/14}	{1/1837}	{1/192}
{1/22}	{1/24}	{1/34}
{1/360}	{1/36}	{1/4000}
{1/6400}	{1/64}	{1/72}
{1/746}	{2/10}	{2/5}
{25/1000}	{25/100}	{3/2}
{3/8}	{4/5}	{1/6}
{1/8}	{2/3}	{1/100}
{1/12}	{2/4}	{4/4}
{5/8}	{6/8}	{1/10}
{1/16}	{1/1000}	{1/20}
{1/3}	{3/4}	{1/4}
{1/2}		