THOMAS M. PAIKEDAY

THE AMERICAN HERITAGE INTERMEDIATE CORPUS

Sample studies of the five-million-word American Heritage Intermediate Corpus, the largest yet for English, show the inadequacies of computerized word counts for lexicographical purposes. Out of about 87,000 word types listed in the *American Heritage Word Frequency Book* (by JOHN B. CARROLL *et alii*, Houghton Mifflin-American Heritage, New York, 1971), only about 40 % seem distinctive lexical items. On the other hand, 22 % of the vocabulary entries required by the 50,000-entry dictionary for which the Corpus was prepared did not occur even once among the five million word tokens.

We first examined three randomly selected portions of the listing in the *American Heritage Word Frequency Book* (*AHWFB*) side by side with the corresponding entry lists of the *American Heritage School Dictionary* (*AHSD*), the *Thorndike-Barnhart Intermediate Dictionary* (*TBID*), the Merriam-Webster *New Students Dictionary* (*WNSD*), the *Holt Intermediate Dictionary* (*HID*), and the *Random House School Dictionary* (*RHSD*). Each of the three portions checked in *AHWFB* was based on a consolidated list of 100 entry words (300 in all), made up of the vocabulary entries in the dictionaries: (1) AFTER to AHAB, (2) CABBAGE to CALEB, and (3) LAKE to LANOLIN.

Our aim was to find out how useful the Corpus was in assessing the vocabulary of Grades 3-9 and in adding new words to the lexicon of that level as presented in the dictionaries.

Here are some of our findings:

There are approximately 110 word types in each sample of the main *AHWFB* listings (once-occurring entries listed at bottom of pages were omitted) corresponding to each 100-entry-word portion from the dictionaries.

Of the 110 word types, 36 % are lexically undistinctive items inadmissible as vocabulary entries in dictionaries. Such are: (a) hyphenated loose compounds, e.g. *After-Shaving*, *agreedon*, and *air-breathing*; (b) solid forms of hyphenated compounds; (c) spellings with undistinctive initial capital instead of lower-case letter, such as *After*, *Against*, *Age*,

and *Ahead*; (d) regular inflected forms of nouns, adjectives, verbs, and adverbs; (c) unusual word types, numbers, and other such " graphic types ".

Over 13 % are lexical and non-lexical items inadmissible in any of the five dictionaries, including *AHSD*. The complete listing of such items is given below:

afterhold	cabildo	lakeside
Agard	cabman	Lamar
Agatha	Cabral	Lambert
Agba	Cabrillo	Lambu
Agelaos	Caccini	lampshade
Aggy	Cadillac*	Lamson
ag'in	Cadore	Lanciotti
Agnes	Caecilian	landcrab
Agootuk	Caernarvon	Landry
Agramonte	Cahokia	Landshort
	Caipira	Langanes
	calamus	Langelinie
	Calchas	Langewiesche
	Calder	Langley
	Caldwell	Langmier
	Cale	Langston
	Caleb	

In addition to the above, *AHSD* had to discard also the following items (5 %):

Agamemnon	Cadi	lamplight
Agassiz	Cadiz	Lamplighter
Agincourt	Calais	Lancaster
Agnew	Calderon	Lancastrian
		landfall
		landless
		landmass

Summing up, according to the entry criteria used by *AHSD* alone, 175 (54 %) out of a total word-type listing of 322 items in the *AHWFB* samples had to be discarded.

---

* N.B. There seems no reason why high-frequency items like *Cadillac* (72) and *Ford* (109) should not be entered in dictionaries, especially when they occur in sentences of this sort: « A Cadillac crashed into a Ford », in which *Cadillac* and *Ford* are used in a generic manner, not as proper names.

In other words, only 46 % of the word types sampled from AHWFB were found to have been useful for AHSD, and these were already in the other dictionaries.

On the other hand, out of the 300 entry words in the three consolidated lists based on dictionary entries, only 235 (about 78 %) were in *AHWFB*. The other 65 lexical items (22 %) which the dictionaries collectively had, but did not occur even once in the Intermediate Corpus, are:

afterbirth
aftercare
afterdeck
afterglow
aftertaste
ageratum
agglomerate
agglomeration
agglutinate
agglutination
agglutinin
aggrandize
aglitter
agnosticism
agog
agora
Agra
agt.
cabby
cabdriver
cabriolet
cacique
cacomistle
cacophonous
caddish
Caddo
Caddoan
Cadette
cadge
Caduceus
Caesarea
Caesarean
caesium
caesura
caftan
cagey
cahoots
caitiff
cajolery
calamine
calash
calcify
calcine
calculable
Laker
lam
lamasery
lambda
lambent
lambkin
lambskin
lamentably
lampoon
lampoonery
Lanai
Lancashire
lancelet
Lancelot
lancer
lancewood
landau
landsman
landwards
lankily
lanolin

In other words, 22 % of required dictionary entries are not in the sample of five-million word tokens (including 86,741 word types) drawn from 1045 texts of grade 3-9 level fed into American Heritage computers.

New words entered in *AHSD* that are not in the 300-word entry list based on the other dictionaries are the following two which seem rare birds occurring not even once in either the Intermediate Corpus or the Brown University corpus:

CACOMISTLE: A racoonlike animal of southwestern North America, having a long black-banded tail;

LAKERS: The national Basketball Association team from Los Angeles.

Another study of the American Heritage Intermediate Corpus that we did was an analysis of its lexical components.

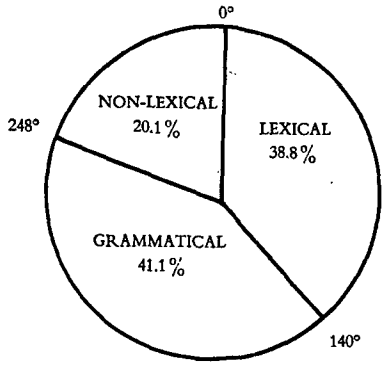
We listed all the word types (including those of single occurrence) in the three portions used in our study of the usefulness of the corpus for lexicographical purposes, viz. (1) AFTER to AHAB, (2) CABBAGE to CALEB, and (3) LAKE to LANOLIN. Then we grouped them under the following headings:

a) *Lexical items*: Under this head we included all word types having lexical meaning, including also names of persons and places applicable to more than one individual. Thus *Caesar* was included but not *Agnew*. Spelling variants such as *Caddy/caddie/Caddie* were counted as one, but irregular inflectional forms were counted separately. *Cadillac* was included for obvious reasons.

b) *Grammatical items*: In this we counted all word types based on lexical items, such as spelling variants, regular inflectional forms, compounds and nonce forms having no more meaning than those of their parts, such as *landloving, land-management*, and *land-reform*, etc.

c) *Non-lexical items*: In this we included unusual word types to which we could not assign a meaning, and unique persons and places. A word type such as *ah-h-h* was counted as a lexical item, but *lan'* and *Aggie* were considered non-lexical.

The following chart indicates the components of the lexicon, according to our sample study, of the Grade 3-9 vocabulary presented in the *AHI* Corpus. It must be mentioned here that non-lexical per-



sons and places of the type usually entered in dictionaries amounted to just over 1 % of the word types we sampled.

Our studies seem to throw much light on the relations between frequency word counts and dictionaries. Dr. Carroll, the senior author of the *America Heritage Word Frequency Book*, has said: " My mathematical analysis of the problem of sampling indicates that a truly enormous corpus, say on the order of 500 million words, would be required if the resulting list were to be expected to include all or nearly all items required for a dictionary of respectable size. " (*Research in the Teaching of English*, Fall 1972, N.C.T.E., Urbana, Illinois, pp. 236-237).