STORAGE, RETRIEVAL, AND EDITING OF INFORMATION FOR A DICTIONARY

by

Richard L. Venezky

June 1967

Technical Report #7

STORAGE, RETRIEVAL, AND EDITING OF INFORMATION FOR A DICTIONARY

by

Richard L. Venezky
University of Wisconsin
15 June 1967

Introduction

The purpose of this paper is to describe the progress and future plans for the automation of the Dictionary of American Regional English (D.A.R.E.).[1] Although most of the problems which are discussed here have already been tackled by others -- optical scanning, file maintenance, retrieval of semantic information, on-line editing -- the combination of these items to produce a dictionary has never been attempted before.[2] While parts of dictionary production have been computer-aided, no lexicographer has yet tried to automate, to a high degree, the storage, retrieval, and editing of information for a large dictionary.[3] Although the discussion that follows concerns only the techniques which have been developed for the current dictionary, these processes are equally applicable to the compilation not only of other dictionaries, but of any publications which are derived from vast, complex stores of data.

The Dictionary of American Regional English

The D.A.R.E. will differ from an ordinary commercial dictionary chiefly in that the D.A.R.E. will concern itself entirely with words which do not have national currency, while most commercial dictionaries concern themselves primarily with words which do.[4] Every region has its special lexicon of words and phrases; what is called a creek in the Midwest becomes a branch in the South, a brook in the New England states, and a kill in certain communities along the Hudson River. Children in the Midwest play a stick game which they call knick-knock. The same game played in Upstate New York is called pee-wee, while in Texas it is cricket. These are samples of the information which the present project will attempt to collect through methods which will be described presently. The primary goal, however, is to gather not only a large store of regionalisms, but also to obtain data on the geographical area in which a particular term is used, the type of people who use it, and what it means. In the five-year period allotted to data collection, the project expects to collect approximately five-million records, each record being a piece of information about a word or phrase which occurs or has occurred in regional speech or in regional literature.

The final product will be a dictionary which gives, besides the part of speech, pronunciation, and definition for each entry, the area

of the country where the word occurs, the type of people who use it (relative age, educational background, extent of social contact), and the currency of the word, that is, whether it is used now, or if not, when it was used. Included will be topical indexes which list all of the dictionary entries for particular categories (e.g., weather, children's games) and maps which show the geographical distribution of selected terms. One map, for example, might show where the terms creek, branch, and kill are used. While this dictionary will probably not become a best seller, it will serve a number of scholarly needs. It will show, for example, the differences in speech of urban dwellers vs. city dwellers, of older, less-educated people vs. younger, more-educated, and of Western speech vs. Eastern speech. In addition, it will be a ready reference for anyone who wants to find the meaning and geographical distribution of regional terms.

The Central File

The automation of this project is based on the establishment of a central file into which new records are entered and from which information, as desired, is retrieved. A large volume of data - probably equivalent to 200 million characters - will be input to the file and stored so that separate parts of a record can be retrieved at any time. In addition, editing will be done on-line with a display scope and keyboard. Information from the

central file will be retrieved, displayed on the scope, edited, and then stored in a new file which will go directly to a printer for publication.

Each record entered into the central file is composed of overtly marked items. The canonical form for a stored record, hereafter called an element, is composed of thirteen separate items.

Element Structure

1.	code	An integer, used as a unique identifier for each element in the file.
2.	headword	The word or phrase in standard American English spelling which the element describes.
3.	pronunciation	Pronunciation in coded phonetic terms.
4.	grammatical class	Part of speech.
5.	variant spelling	The spelling of the actual response if item 2 is a normalized response.
6.	geographic location	City (subitem 1) and state (subitem 2).
7.	informant type	Education (subitem 1), occupation (subitem 2), and age group (subitem 3).
8.	source	Origin of information. The first subitem contains the source class, the second, the designation within the class.
9.	sense type	Semantic classifiers.
10.	definitions	The most common English equivalent, if one exists, (subitem 1), and an expanded definition (subitem 2).

11.	citation	Utterance in which the headword occurs.
12.	usage data and informant date of birth	Point in time at which the headword occurred (subitem 1), date of birth of informant (subitem 2).
13.	notes	Information which does not fit into any other item.

In item 7, an informant's education is classed as:

COL	at least two years beyond high school
HS	at least two years of high school
GR	at least two years of grade school
N	None
X	Unknown

Occupation is classed as:

BUS	Business
FAR	Farming
HOU	Housewife
OFF	Office employee
PRO	Professional
SKL	Skilled laborer
UNS	Unskilled laborer

Age groups are O (61 or older), M (30-60), and Y (up to 29).

Subitem one of source (item 8) could be any of the following:

W	Wisconsin English Language Survey (a study of Wisconsin speech, carried out in 1950).
D	D.A.R.E. questionnaire (see p. 8).
T	D.A.R.E. tape recordings (see p. 8).
R	Reading program (see p. 9).
L	Linguistic Atlas (a major study of American speech, started around 1930 and nearing completion now).

Subitem two for W, D, or L is the section and number of the question on the questionnaire; for R a short title, and for T, the informants code.

All elements have a code, headword, and source. The other items that an element contains vary with the input data from which the element was formed.

Phrases are entered into the file under each significant word in the phrase (non-significant words are function words like a, and, the). Item 2 for a phrase contains two subitems: a significant word from the phrase and the phrase itself. For example, "Snake in the grass" would appear in two elements with the following item 2 entries:

element 1	snake * snake in the grass
element 2	grass * snake in the grass

Only the first element, however, would contain full entries for the remaining items. All succeeding elements would contain a code, headword, source, and then in item 13 (notes) a cross reference to the first element for that phrase. A group of elements from the D.A.R.E. file are shown in Figure 1.

Figure 1 here

Items are either unary or multiple, depending upon whether they are composed of a single piece of information (which may itself be composed of any number of characters or words) or of two or more separate pieces of information. Code, for example, is a unary item, consisting of an integer, while informant type is a multiple item, consisting of age, education and occupation. In addition, items are classed as numeric, alphanumeric, or coded, depending upon the form of the information they contain.

The retrieval system which operates on the central file retrieves information on the basis of questions asked about items within an element. For example, someone might want to retrieve the headword and pronunciation of every element having a certain geographic location and a certain informant type. In addition to retrieving information, one can also update the file, change the canonical element structure, validate the contents of the file, sort retrieved information, and print reports and word distribution maps.

The use of the central file involves three stages. Stage one is data input -- reduction of data from its initial form to a form acceptable by the central file. Stage two is storage and retrieval -- entering data into the central file and retrieving information from it. Stage three is editing -- retrieving and displaying information, altering it, and storing it in a new file.

Stage I -- Data Input

The raw data for the Dictionary come from three major sources: field workers, a reading program, and dialect publications.

Field workers

During the data collection period field workers travel throughout the U. S. in camper wagons, recording the speech of the people they interview. The communities they visit and the people they interview are selected through settlement history studies. The number of people interviewed in each state depends upon the state population; we intend to interview approximately 1,000 persons in all, being careful to maintain certain percentages of age, education, and occupation in the people who are interviewed.

Field workers return to the project for each interview an hour tape recording of conversation with the person being interviewed and a filled-out questionnaire. The questionnaire contains approximately 1,500

questions which the interviewer asks (in an established form which will ensure comparability in the answers). Some typical questions are:

1. If a day is very hot, you say it's a _____.

   (scorcher, sizzler, hotter than Dutch love, a day for a carp roast, etc.).

2. What do you call the time in the early morning when the sun comes into sight?

   (daybreak, sunrise, sunup).

Most of the time the field worker records the response in traditional orthography, but where pronunciation is significant, phonetic spelling is used. To date, several hundred questionnaires have been completed and pre-edited.

Reading program

Novels, travelogs, magazine and newspaper articles, and diaries are potential sources of regional vocabulary. To help tap these sources, members of the American Dialect Society and other interested persons have volunteered to read documents, marking passages which contain words of regional usage. Some of the reading materials are selected by the project through research on regional literature, and some are selected by readers who are especially familiar with the writings of their region. Readers underline each word or phrase which is of regional interest, enclosing in brackets the relevant context in which it occurs.

The documents are then returned to the D.A.R.E. office for pre-editing and typing. An editor reads each document, attempting to ascertain the authenticity of the regionalisms. If this test is passed successfully, margin slips are pasted on each page which has an underlined item. An editor then marks on the margin slip the following information for each citation:

1. The standard spelling of underlined words which are not spelled with traditional orthography. (Some authors, for example, attempt to spell regionalisms phonetically.)
2. Code number of speaker. Each person in the document who speaks is given a code number.
3. Sense type. One or more descriptors are assigned to each underlined word.
4. Date. A date is assigned to each document on the basis of the period in which the majority of the action takes place. If a citation has been placed by the author in a period different from this assigned date, a new date is entered on the margin slip by the editor.

Finally, a heading is added to each document, identifying the title, author, date, geographical location, reader of the material, and identity of characters created by the author.

Dialect publications

The American Dialect Society has published in Dialect Notes and PADS (Publications of the American Dialect Society) information on approximately 100,000 words or phrases of regional interest. A typical entry is shown below.

> generals, n. Some of the expenses of a fishing voyage are divided equally among the crew, such as food and the cook's wages; these are called small generals. Other expenses are divided in proportion to each man's catch of fish, as bait, salt, and barrels; these are called great generals or big generals.[5]

The dictionary staff processes the data from these sources as follows:

1. An editor reads each record for clarity and accuracy and adds item designators and sometimes whole items directly to the raw data records.
2. A secretary types each record for scanner input, using a standard typewriter fitted with an ASA scanner font golf-ball.
3. Typed pages are transferred to magnetic tape by an optical character reader. (A limited amount of format conversion is done during scanning.)
4. A format routine converts scanner output to element form for entry in the central data file.

Input records are typed on preprinted fanfold forms which were designed to facilitate both typing and scanning. A form number at the top of each scanner page indicates which scanner control subroutine is to be called for scanning that form. The optical character reader (the CDC 915) reads standard ASA scanner font: 26 alphabetic characters (all upper case), 10 digits, and 25 punctuation and special characters. The special characters are used for program control, e.g., to indicate characters or lines to delete, lines to skip, and special characters to be inserted by the scanner at various points in the reading of a document. A scanner input page is shown in Figure 2.

Figure 2 here

Scanning, in preference to keypunching, not only reduces total data prepraration costs, but also eliminates a number of punched-card imposed burdens. All material is typed for scanner input in the project offices by typists who work for the dictionary project, rather than by physically and mentally remote employees of a private keypunch firm. Changes in data preparation techniques are made on the spot, rather than through an intermediate agent. In addition, pages for scanning are easier to store than punched cards; they weigh less and occupy less space than punched cards with a comparable quantity of information, and are considerably more useful as reference materials. Since almost eighty percent

of the cost of placing material on magnetic tape is absorbed either by the scanner typing or by keypunching, we expect to save a considerable amount of money over keypunching, which costs approximately three times as much as typing in this area. (The operation of the scanner itself, which represents about twenty percent of the preparation costs, is a little less than twice the cost of operating a card reader.)

Stage II - Storage and Retrieval

Information retrieval system

An information storage/retrieval system which satisfies the requirements described above has been implemented by Control Data for the 3400 and 3600 computers, and has been adapted to run under the drum monitor system for the 3600 at the University of Wisconsin.[6] This system, called INFOL, stores data in variable-length elements. Each element is composed of a set of variable-length, but overtly bounded items. The item, or an overtly bounded subitem, is the smallest unit which can be manipulated under this system. (See page 4 for a description of the element used for storing dictionary data.) Thus, interrogations, extractions, and updates can be based upon items or subitems, but not upon units smaller than these. During file establishment INFOL creates a basefile which stores the programmer supplied description of the basic

element and bookkeeping information compiled by the system on the actual elements which are entered into the file. Basefile becomes the first physical record in the information file; it provides the system with control information during interrogation, extraction, updating, and validation phases. Interrogations and corresponding extractions are described in a reasonably flexible grammar. In addition, INFOL has its own sort/merge system and four different report generators. (Two report generators produce formatted output for input to other computer programs, while the other two produce printer output.)

It is the flexibility of this retrieval system which distinguishes the D.A.R.E. system from most other publication schemes. Within the pale of traditional lexicography, an editor is limited by a fixed filing system, within which record classifications, other than by alphabetic sequence and sometimes by subject, are forever entombed. Even in some of the recently publicized computer-aided schemes, only a small number of fixed fields can be designated for sorting -- all other relationships being lost. While these retrieval abilities may not be needed in the compilation of a commercial dictionary which can draw upon its predecessors and its competition, they are essential for the D.A.R.E. which must obtain the majority of its content through its own research activities.[7]

In the D.A.R.E. system elements can be retrieved on the basis

of the value of any item, or upon the basis of various logical combinations of values of different items. In addition, the user can specify what items are to be extracted from those elements which satisfy a particular interrogation. For example, the following coding requests INFOL to find all elements in which the informant's occupation was not Business. For items which meet this qualification, the headword, informant occupation, and headword definition are to be extracted and printed in a standard format. The results from a test file are shown in Figure 3.

```
INTERROGATIONS
    *7* SUB-ITEM 2 NE BUS
EXTRACTIONS
    *2* REPORT
    *7* SUB-ITEM 2 REPORT
    *10* REPORT
```

Figure 3 here

Once parts of elements are extracted, any item, or subitem can be designated as a key for sorting.

The statements shown below are a request for INFOL to find every element for which the informant had no more than a high school education and was classed as old. The location, occupation, and date of birth

from each such element were extracted, and then sorted into chronological order on date of birth. The results are shown in Figure 4.

```
INTERROGATIONS
    *7* SUB-ITEM 1 EQ HS OR GR OR N AND SUB-ITEM 3 EQ O
EXTRACTIONS
    *6* REPORT
    *7* SUB-ITEM 2 REPORT
    *12* SUB-ITEM 2 SORTKEY 1
```

Figure 4 here

It is also possible to obtain at any time a complete alphabetized list of headwords and head phrases, along with a count of the number of times each occurs in the file. The coding for this request is:

```
INTERROGATIONS
    *2* EXISTS
EXTRACTIONS
    *2* INVERT
```

Part of the results are shown in Figure 5. The numbers to the right of the words or phrases are the code numbers of the elements from which they were extracted.

Figure 5 here

Retrieval of semantic information

The retrieval of information based upon any items except definition and sense type is relatively uncomplicated. Sense type and definition, however, embody all of the problems that are commonly discussed under the title of indexing, or classification, or the library problem.[8] A general solution of these problems in, for example, the automation of a library, must deal with the universe of all possible meanings and the task of locating data that sit at particular points in the meaning space. Semantic retrieval problems for this dictionary, while vast, are not so astronomical. The D.A.R.E. is concerned with a limited subset of the English lexicon -- a subset which contains a large number of referential terms and a small number of difficult to define abstractions. In addition, it does not have equal interests in all possible connections among the items in our store, but instead can assign weights or priorities to many of the potential links. Thus, we can aim for dense subcategorization in one class while being satisfied with much coarser divisions in another.

The major connections among non-identical terms are, in the expected order of descending importance to our project, the following:

1. Definitional identity. Two or more terms often refer to the same object, emotional state, or process. Tote, pail and bucket, for example, are different names for the same

physical object. Connections of this type form the heart of an analysis of regional variations in speech by permitting the direct comparison of regional vocabulary.

2. Sense class identity. Different terms often share a sense dimension, such as naming domestic animals or farm implements, or drag-racing procedures.
3. Etymological identity. Different terms may share a feature of etymology, such as being spelling pronunciations or French loan words. Such classifications are especially valuable for studying how regional vocabulary is formed.
4. Functional identity. Two or more terms may share a functional feature, e.g., being euphemisms or exaggerations.

Except for sense class identity, which could be pursued into complete chaos, all of these connections could be discovered if enough editors worked enough years with the data (or if enough monkeys poked randomly at classification keys for a sufficient infinity). The real challenge, however, is to construct more practical discovery procedures for the required classifications. Part of this was done in the design of the field work questionnaires and in the design of the basic element for storage. Definitional identity, for example, can be retrieved from item 10(source -- see page 4) for headwords obtained by field workers.

Source identifies a particular query on the questionnaire so that it is possible, by locating the question that attempts to elicit a description of some object, process, or state, to then retrieve all of the responses that fit that description.

In some elements definitional identity can be retrieved from item 8 (definition), although this item is not totally reliable since semantically identical definitions might be verbalized differently (cf. "a frying pan": "a pan for frying"). Some sense class identities are either implicit in the source or are coded into the raw input data. Further sense class coding using, for example, a key word system, can be done on update passes once a functional classification scheme is derived. Structural and functional identities must be added during initial input or during updating.

Two techniques are being employed for discovering connections within the central file. One is by secondary associations. During one pass all elements which have already been classed in a particular category are retrieved. A second retrieval pass is then made to retrieve all remaining elements which have headwords or definitions which match those of items retrieved on the first pass. Through an analysis of the items obtained on the second pass, connection codes are added to selected elements in the file.

The second technique is a simple key-word-in-context scheme. A concordance is made to the definitions in the file, and from an analysis of the concordance further connections within the file are established. Schemes one and two could be combined by first making a concordance and then using already established connections to form groups of potentially related terms.

State III - Editing

To facilitate editing, we plan to develop an on-line editing system for natural language texts. This system will probably be developed for an inexpensive computer and be based upon general retrieval and list processing sub-systems. If the output from the editing is to be a tape which will drive a linotype machine or some other typesetting device, then the display driver must be capable of forming and coding for output a variety of character styles -- standard, italics, and bold face, to name just a few.

The editor will have a scope-keyboard module where the edited text appears on one side of the scope and unedited text retrieved from the central file appears on the other side. He will request that elements be retrieved and displayed, select portions of elements which are displayed on the raw data side of the scope and insert them in the edited format which is displayed on the other side of the scope, and finally,

transfer edited text to the output tape. The software for this system will include the following:

1. An impoverished retrieval system that retrieves either consecutive elements, or locates non-consecutive ones on the basis of a single item interrogation. No sort/merge routine or report writer will be needed.
2. A list processing system that stores retrieved information for displaying and editing, and performs such primitive list processing functions as adding or deleting items from a list.
3. A scope display system that takes data from the list processing storage and displays it in a requested form.

A simple editing system for an alphanumeric CRT terminal (Control Data 212) has been developed and is now being tested at the University of Wisconsin on the 3600. This system allows file-to-file editing through insert, delete, and replace instructions. The CRT holds twenty lines of fifty characters each, and has a symbol repertoire of 63 symbols. Instructions are entered on the top line, leaving nineteen lines for text. The edit routine operates as a non-resident background routine and therefore consumes little CPU time. The limited symbol repertoire and the small CRT size, however, make this system impractical for production editing. Only with a graphical terminal which has at least 128 different

characters, plus size and orientation variations, can respectable editing be performed.

Conclusions

Automating to the degree that I have outlined above has been relatively expensive, but should lead to major savings in time and overall costs for dictionary production, besides allowing a degree of cross-indexing and categorization that has never been achieved in lexicographical publications. Under the system described above we are able to assemble data in ways which can be done only clumsily, if at all, under the punch-sort or manual procedures. It is possible, for example, to

- learn the coverage at any time of a geographical area, an informant type, a word-class, or a time-period;
- assemble all records obtained from a particular source or during a particular time period; assemble all records which fail to meet certain criteria, such as having a source, or citation or definition;
- convert the pronunciation or any other coded portion of a record from one system to another;
- assemble all records which have the same definition or have a certain percentage of overlap in definitions.

Dictionaries of the future will most certainly be automated, but to what degree and with what advantages over present methodology can not be predicted at present. We expect, however, to be able to answer these questions within the next five years.[9]

FOOTNOTES

1. The work described here is being done at the University of Wisconsin under U. S. Office of Education Grant No. 2935.

2. See, for example, G. Arnovick, et.al., "Information storage and retrieval: analysis of the state-of-the art," Spring Joint Computer Conference 25.537-62 (1964) and Ben-Ami Lipetz, "Information storage and retrieval," Scientific American 215.224-42 (Sept., 1966).

3. A brief, non-technical discussion of a punched card and paper tape system enployed in the publication of a dictionary is described in Laurence Urdang, "The systems designs and devices used to process The Random House Dictionary of the English Language," Computers and the Humanities 1.31-33 (Nov., 1966).

4. For a general discussion of dictionary preparation, see Fred W. Householder and Sal Saporta, eds., Problems in Lexicography. Publication 21 of the Indiana University Research Center in Anthropology, Folklore, and Linguistics (Bloomington, 1962).

5. Dialect Notes 2.425 (1904).

6. See INFOL Reference Manual, Pub. No. 60170300 (Palo Alto, Calif.: Control Data Corporation, 1966).

7. A different approach to the retrieval of dialect materials was proposed by Roger W. Shuy; "An automatic retrieval program for the Linguistic Atlas of the United States and Canada," in Paul Garvin, ed. Computation in Linguistics: a case book. Bloomington:Indiana University Press, 1966.

8. For a discussion of these problems, see C. P. Bourne, Methods of Information Handling. New York: John Wiley and Sons, 1963, chs. 1-3.

9. Several other approaches to complete publication systems have been published recently. See Van Zandt Williams, et.al., "Consideration of a Physics Information System," Physics Today 19.45-53 (Jan., 1966) and P. F. Santarelli, "An Automated Publications Concept," IBM Document TR 00.1263-3 (March 11, 1966).

```
A b     |  |  &&15| 00015

A     THREATENING|
AA    UNSETTLED⋈UNDECIDED|
B     ¥¥ GOING TO BE HOMELY⋈13⋈▽BELGIAN▽|
BB    THREATENING⋈INCLEMENT|
C     THREATENING⋈13⋈0CC|
CC    THREATENING|
D     THREATENING|
DX    UNSETTLED|
E     THREATENING|
EE    THREATENING|
EEX   THREATING|
F     THREATENING|
FF    THREATENING|
H     LOOKS LIKE WE'RE GOING TO HAVE A SPELL OF WEATHER|
HH    THREATENING|
I     THREATENING|
II    STORM IS BREWING⋈WE'RE GOING TO GET SOMETHING|
J     MAKING FOR BAD WEATHER|
```

Figure 2 -- Scanner input page (upper half)

LIST OF ALL ELEMENTS IN THE DARE FILE

CODE	20
HEADWORD	SNAKE * SNAKE IN THE GRASS
PRONUNCIATION	
FORM CLASS	
VARIANT SPELLINGS	
SOURCE LOCATION	MONROE * WI
INFORMANT TYPE	COL * MOD * C
SOURCE CLASS	W * G57
SENSE TYPES	DECEIT * ADAGE
DEFINITION	X * DECEPTIVE SITUATION
CITATION	
USAGE-INF. DATES	1950 * 1975
NOTES	

CODE	21
HEADWORD	GRASS * SNAKE IN THE GRASS
PRONUNCIATION	
FORM CLASS	
VARIANT SPELLINGS	
SOURCE LOCATION	
INFORMANT TYPE	
SOURCE CLASS	W * G57
SENSE TYPES	
DEFINITION	
CITATION	
USAGE-INF. DATES	
NOTES	SR * 20

CODE	22
HEADWORD	HOUSEBOW
PRONUNCIATION	
FORM CLASS	
VARIANT SPELLINGS	
SOURCE LOCATION	READSTOWN * WI
INFORMANT TYPE	HS * FAM * M
SOURCE CLASS	W * G99
SENSE TYPES	BUILDINGS * GOVERNMENT * PROTECTION
DEFINITION	JAIL * A
CITATION	
USAGE-INF. DATES	1950 * 1967
NOTES	

Figure 1 -- Three elements in the D.A.R.E. test file

HEADWORD Will-O-The-Wisp
INFORMANT TYPE SUB-ITEM 2 OFF
DEFINITION A * A SMALL LIGHT THAT DANCES OVER A SWAMP

HEADWORD ELF
INFORMANT TYPE SUB-ITEM 2 OFF
DEFINITION A * FAIRY TALE CREATURE

HEADWORD PIXIE
INFORMANT TYPE SUB-ITEM 2 OFF
DEFINITION A * FAIRY TALE CREATURE

HEADWORD FAIRY
INFORMANT TYPE SUB-ITEM 2 OFF
DEFINITION A * FAIRY TALE CREATURE

HEADWORD BROWNIE
INFORMANT TYPE SUB-ITEM 2 OFF
DEFINITION A * FAIRY TALE CREATURE

HEADWORD PIXIE
INFORMANT TYPE SUB-ITEM 2 OFF
DEFINITION A * FAIRY TALE CREATURE

HEADWORD STALE
INFORMANT TYPE SUB-ITEM 2 SKL
DEFINITION A * OLD JOKE THAT ISNT FUNNY

HEADWORD CHESTNUTS
INFORMANT TYPE SUB-ITEM 2 SKL
DEFINITION A * OLD JOKE THAT ISNT FUNNY

HEADWORD PAUNCH
INFORMANT TYPE SUB-ITEM 2 SKL
DEFINITION A * OVERSIZED STOMACH

HEADWORD BAY * BAY WINDOW
INFORMANT TYPE SUB-ITEM 2 SKL
DEFINITION A * OVERSIZED STOMACH

HEADWORD MILWAUKEE * MILWAUKEE FRONT
INFORMANT TYPE SUB-ITEM 2 SKL
DEFINITION A * OVERSIZED STOMACH

Figure 3 -- Headword, definition and informant occupation for elements in which occupation is not classed as Business

```
SOURCE LOCATION  HARTFORD * WI
INFORMANT TYPE   SUB-ITEM 2 BUSINESS
USAGE-INF. DATES SUB-ITEM 2 1870

SOURCE LOCATION  HARTFORD * WI
INFORMANT TYPE   SUB-ITEM 2 BUSINESS
USAGE-INF. DATES SUB-ITEM 2 1870

SOURCE LOCATION  ANTIGO * WI
INFORMANT TYPE   SUB-ITEM 2 HOUSEWIFE
USAGE-INF. DATES SUB-ITEM 2 1895

SOURCE LOCATION  ANTIGO * WI
INFORMANT TYPE   SUB-ITEM 2 HOUSEWIFE
USAGE-INF. DATES SUB-ITEM 2 1895

SOURCE LOCATION  ANTIGO * WI
INFORMANT TYPE   SUB-ITEM 2 HOUSEWIFE
USAGE-INF. DATES SUB-ITEM 2 1895

SOURCE LOCATION  BEAVER DAM * WI
INFORMANT TYPE   SUB-ITEM 2 HOUSEWIFE
USAGE-INF. DATES SUB-ITEM 2 1900

SOURCE LOCATION  WISCONSIN RAPIDS * WI
INFORMANT TYPE   SUB-ITEM 2 HOUSEWIFE
USAGE-INF. DATES SUB-ITEM 2 1904
```

Figure 4 -- Location, occupation, and date of birth of informants who have no more than a high school education and are classed as old. Records are listed in chronological order according to date of birth.

1	BAY	39
1	BAY WINDOW	39
1	BEER BELLY	42
1	BELLY	42
1	BROWNIE	32
1	CHESTNUTS	37
1	CROWBAR	24
2	CROWBAR HOTEL	24 25
1	ELF	29
1	FAIRY	31
1	FENCE	9
1	FRONT	44
2	GOOD MILKER	12 13
1	GOOD	12
1	GOT SKINNED	17
1	GRASS	21

Figure 5 -- Alphabetized list of headwords and headphrases