CENSUS BUREAU EXPERIENCE WITH AUTOMATED INDUSTRY AND OCCUPATION CODING

Martin V. Appel, Eli Hellerman, Bureau of the Census

1. INTRODUCTION

All data collection organizations encounter applications which require the analysis of natural language text for the purpose of selecting a numeric code from a classification system. Codes are more amenable to computer manipulation. Discrete categories are needed for tabular presentation and aggregation to higher levels. In a clerical coding operation the coder refers to a list or dictionary of concepts which defines the translation. This clerical assignment of codes is subject to human error and subjective judgment which are not consistent when tracked over time. An incomplete list imposes a burden on the coders. Recruiting, training, and managing large numbers of coders imposes a burden on management.

Industry and Occupation (I&O) data are collected in the Census Bureau's Population Census and most continuing household surveys, either as primary characteristics as in the Current Population Survey (CPS), or as explanatory variables in surveys such as the National Crime Survey (NCS), Annual Housing Survey (AHS), Consumer Expenditure Survey (CE), and others. The clerical coding and verification costs associated with classifying I&O natural language descriptions is about $1.2 million per year for all demographic surveys. The 1980 Population Census spent approximately $7.7 million on this activity.

To help in understanding the clerical coding process, consider the following example, with responses to I&O questions capitalized and underlined.

INDUSTRY QUESTIONS

- For whom did this person work:
  POST
- What kind of business or industry was it:
  NEWSPAPER PUBLISHER
- Is it mainly (MFG, WHSLE, RETAIL, OTHER): OTHER

OCCUPATION QUESTIONS

- What kind of work was this person doing:
  SALES PERSON
- What was this person's most important duties:
  SELLS ADVERTISING
- Was this person an employee for private company, Government employee, self-employed or working without pay:
  PRIVATE

A coding clerk looks up the industry response "Newspaper Publisher" in a dictionary and retrieves its associated code "171". The clerk then looks up the occupation response "Salesperson" and sees that for "Salesperson, Advertising" the code is "256". Note that the coder recognized that "Sales Person" should be one word and combined "Salesperson" and "Advertising" from two separate phrases.

The Census Bureau has experimented with automated I&O coding for several years with the objective of having the computer assign classification codes based on the natural language responses. It was not expected that such a system would reduce coding costs (unless an Optical Character Reading, (OCR) capability became available), since the keying costs alone for textual responses would roughly equal clerical coding costs. However, it was believed that the opportunities for better coding quality, improved consistency, systematic ways to strengthen the coding structure and procedures, and the reduced managerial burden from substituting keyers for coders in large numbers would justify the added expense. Keyers are generally available, need little special training, and are engaged in a familiar, controllable operation; I&O coders require substantial special training and a long learning process. The clerical coding operation also requires strong technical support and is often difficult to control. Now with the advent of the video scan concept (described in Section 4) and with most demographic surveys already keying other data items, it is believed by the authors that significant cost savings can be realized through automated coding (Hellerman 1978).

This paper presents a brief history of the Bureau's experience in developing an automated system (Section 2), an overview of the concepts used in the latest attempt at computerized I&O coding (Section 3), data entry problems and solutions (Section 4), potential usage of the current system (Section 5), and the status of current research (Section 6).

2. HISTORY

An automated coding system was developed prior to the 1967 Economic Censuses, which used two basic algorithms, and assigned industry codes (as defined by the 1967 Standard Industrial Classification (SIC) Manual) to natural language questionnaire responses provided by small business establishments. The first algorithm constructed a reference list for industry codes on the basis of historical response/code sets; the second coded new responses from this reference list. The SIC industry coding structure, used for establishments, differs from the I&O industry coding structure, used for individuals.

This initial effort, called the O'Reagan algorithm (O'Reagan 1972), and the two which followed in sequence placed great emphasis on the observed frequencies of responses and associated codes. More recent attempts have concentrated on computerizing the clerical coding procedure.

For an example of how that first algorithm proceeded, let us assume an eight-word, three-code sub-universe and a response language in which each capital letter is a proxy for some unique word. Assume further that the entire experience of all eight words, all past responses which used any of them, and the associated clerically assigned codes are available as computer records, in the following historical file:

Record #	Response	Code
1	AB	1
2	ABCK	1
3	KWQ	1
4	BQ	2
5	WR	2
6	QM	3
7	MM	3

Word A only occurs in responses coded to category 1. If A was "physician", Code 1 of the example might correspond to Code 8011 in the SIC manual. Words K and C likewise are associated only with Code 1, and words R and M with Codes 2 and 3 respectively (MM is a "wires wires" type of response). The primary reference list constructed from this history file would then be:

Record #s	Presence of word _	Implies Code _
1,2	A	1
2,3	K	1
3	C	1
5	R	2
6,7	M	3

Eliminate these classifier words from the original file and there remain:

Record #	Response	Code
1	B	1
2	B	1
3	WQ	1
4	BQ	2
5	W	2
6	Q	3

B together with Q identifies Code 2, and B without Q is always Code 1. For instance, "doctor veterinary" might be always coded 0742 but "doctor" without "veterinary" is again coded to 8011. Using set notation where $\bar{Q}$ means "anything that is not Q", the new part of the reference list is:

Record #s	Presence of words _	Implies Code _
1,2	$B\bar{Q}$	1
3	WQ	1
4	BQ	2
5	$W\bar{Q}$	2
6	$Q\,\bar{W}\,\bar{B}$	3

Combine this secondary list with the primary reference list and one has an adequate computer dictionary to handle any new response which is similar to past experience. Analysis showed that word strings longer than four seemed to offer no benefit; string length two was most profitable. The algorithm is described more mathematically in one of the references (O'Reagan 1972).

When large samples of coded responses are used to develop the reference list, few actual word strings satisfy the criterion "word string i unequivocally implies code j", partly because of inconsistent clerical codes in the history file. In practice, that criterion is replaced by "in at least T_1% of all historical cases of word string i, it occurs with code j". Use of this parameter permits trading some risk of miscoding for increased ability to code. Use of another parameter T_2 permits coding of incomplete responses. If 86 percent of the experience "nursery" is accompanied by "school" and coded to 8351, and 14 percent of the time it is joined by "tree" and coded to 0783, a T_2 threshold of 85 percent built into the coding algorithm would permit the unaccompanied word "nursery" to be coded to 8351. A third system parameter specified the minumum number of times, N, that a word string must appear in the history file to be eligible for the reference list. This constraint would prevent "Betty's" in the unique record "Betty's Beauty Shoppe" from becoming a unique classifier.

In its best test performance, this system coded 86 percent of incoming responses, with 96 percent accuracy. Under production coding conditions, the system coded 79 percent of 788,000 responses, 92 percent of those correctly. Special clerks had difficulty with the remainder. Construction of the reference file from 10,000 records needed 20 minutes from a UNIVAC 1107. The coding pass ran 2700 records per central processor minute.

The Corbett algorithm (Corbett 1972), devised in 1972 but not tested until several years later, used 1970 Population Census records and the code structure from the 1970 Classified Index of Industries and Occupations. As with the previous technique, a word string must appear some minimal number of times, N, in the history file, and the word or string must be coded to the same category in at least T_1% of its occurrences to be a classifier. However, the system made no provision for a T_2 parameter. The aim was to find the minimal set of word strings that would define the code, logically equivalent to finding the fewest switches in a Boolean switching function.

Consider, for example, this hypothetical history file which includes the experience of all records containing any of the words A, B, C, or D:

	Word				
Record #	A	B	C	D	Code
1	1	0	0	1	1
2	0	1	1	0	2
3	0	1	0	1	3
4	0	1	1	1	3

Zero in the word matrix signifies absence of the word and 1 signifies presence.

It can be shown by Boolean analysis that two tests are sufficient to code any new record which matches this history file—knowledge of whether word A occurred, plus knowledge of whether word D occurred. Note that this algorithm differs from the one above in its handling of the T_2 or incompleteness condition. When a classifier word or string is identified in the historical file, all other words in all records containing that classifier are eliminated from further examination. The entire algorithm is more fully described in Corbett (1972).

Testing in 1975 and 1976, on 1970 Population Census records, yielded results inferior to those obtained by the O'Reagan algorithm, which coded to industry 72 percent of the records with 88 percent accuracy. The Corbett method was also slower.

From these earlier efforts evolved the Information Measure Processor (IMP) (O'Reagan 1972). It was an attempt to make explicit use of information theory concepts and notations. For every word/code pair in the historical record file a weight was computed and stored in a matrix W, where i was the word index and j the code index.

$$W(i,j) = \log_2 \frac{P(j\,|\,i)}{P(j)}$$

The mutual information (MI) measure was calculated for each code by a summation across all words in the code:

$$MI(j) = \sum_{\text{all }(i,j)\text{ in }j} P(i\,|\,j) \cdot \log_2 \frac{P(j\,|\,i)}{P(j)}$$

To code an incoming record, each response word was searched in the matrix W. For words found, the appropriate weights were added (for each code). The sum for each code j was then divided by MI(j) and that code with the highest result was the code assigned.

Restrictions on this simplest scheme were tested. The code with the maximum score was required to exceed a threshold and also to exceed its nearest competitor by some specified percentage. Utilizing the same files on which the earlier methods were tested, IMP could code the entire input with 77 percent accuracy or code nearly half the input with virtually no errors, depending on the restrictions. Subjectively, perhaps the best balance seemed to be a 73 percent coding rate with 87 percent of those codes correct. The history file was processed at nearly 10,000 records per minute, but coding speed was only 500.

Major and minor variations on all these schemes were tried to the extent that resources permitted. Taking account of word order and syntax produced no benefit. Efforts to synonymize, truncate, or compress (including SOUNDEX) provided ambiguous results. A cascade of approaches, e.g., passing a record on to IMP if Corbett failed to code, was considered but never tested.

By late 1976 the Census Bureau reached some major conclusions:

1) The code structures themselves were not built according to the principles of mathematical taxonomy, and coding procedures left a residue of cases for which the "correctness" of coding could not be determined readily. Also, since in earlier tests all differences between computer-assigned and clerically assigned codes were considered errors in the former, early estimates of system accuracy rate are conservative, and slightly fuzzy.

2) The total cost of computerized coding was roughly equivalent to that of clerical coding unless the history file and the responses to be coded were in computer readable form at no cost.

3) The attainable coding rates and accuracy rates (for a batch configuration) were comparable to those attained by newly trained coders but inferior to the performance of highly experienced coders. In other words, the computerized systems could code well the common and simple responses, but were inadequate for rare or equivocal cases.

4) Unless a relatively few response/code sets cover a large portion of the population (Pareto principle), a computerized approach could well become unwieldy. Over half of all records can be covered by 16 percent of the industry codes.

5) Certain response fields contained greater information than others.

In light of these judgments, a special committee, established to determine the feasibility of an automated I&O coding system for the 1980 census, decided that no system achievable by 1980 appeared practicable for census processing.

3. HELLERMAN I&O CODING SYSTEM (HIOCS)

Automated I&O coding research took a significant change in direction in late 1976. It no longer used a sample of questionnaire responses in the development of the dictionary of phrase/code sets, but relied on the "coding manual" to provide the patterns for matching and the numerical weights for scoring. During this developmental period we studied the coder's instructions, attended coder training and observed production coding, all for the purpose of developing a computer emulation of the coding process. It has been and continues to be an iterative process of research, testing, evaluation and enhancement. Concepts have been adapted from other methodologies or previous research. Sometimes the adaptation has been applied in a non-traditional manner. We will indicate where this has occurred.

HIOCS (Hellerman 1982) is the latest in a progression of research coding systems. It can be categorized as being in the spirit of artificial intelligence, in that it attempts to simulate in the computer the processes that a production coding clerk employs: such things as recognizing "meaningful" words, misspellings, synonyms, and abbreviations. HIOCS also arrives at a classification through inference when the "dictionary" of phrases and codes is deficient.

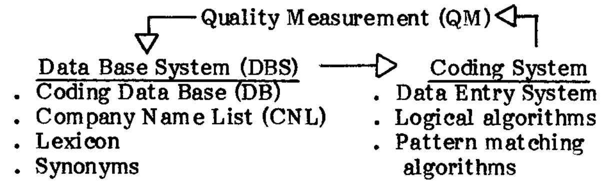
HIOCS consists of three major subsystems.

1) The Data Base System (DBS) establishes and maintains a centralized data base of both files and arrays. It consists of I&O descriptor phrases, company names, a synonyms list and a lexicon of all words in the descriptors.

2) The Coding System incorporates those functions necessary to enter and code a respondent's reply.

3) The Quality Measurement (QM) System measures the accuracy of codes assigned by HIOCS. The system selects a sample of incoming responses for clerical coding and compares the results to computer generated codes. This process detects special or new coding situations requiring improved algorithms or data base changes.

The subsystems are linked logically as shown:



Quality Measurement (QM)

Data Base System (DBS)
. Coding Data Base (DB)
. Company Name List (CNL)
. Lexicon
. Synonyms

Coding System
. Data Entry System
. Logical algorithms
. Pattern matching algorithms

The coding procedure begins by reducing the respondent's words to a "standard" representation. Next, it looks for key words which imply that logical analysis of the responses may yield a code or it matches the respondent's reply to a question to phrases contained in the coding Data Base (DB). The DB, originally the same as the "coding manual" used by clerks, is a computer adaptation and extension of the 1980 Alphabetical Index of Industries and Occupations. The general procedure for DB matching is to find the DB phrase that best fits the respondent's phrases. If there is a direct match, the associated code is assigned. However, direct matches occur infrequently so DB phrases have to be constructed from the words of two or more phrases used by the respondent. Because several phrases may be constructed from the given phrases, it is necessary to "score" the DB phrases on some basis of the closeness of fit.

The general coding procedure is as follows:

Phase 1 (Data Entry and Preparation)

Step 1. Enter responses to each of six I&O questions. The industry section of the questionnaire includes two questions on industry and one on the economic sector. The occupation section contains two questions on occupation and one on class of worker.

Step 2. Develop the "standard" representation for each word in the respondent's reply. Synonym and abbreviation lists are searched for replacements.

The standardization of words is necessary because respondents do not use the phraseology of the coding manual. It also helps eliminate some general uncertainties of spelling (doubling of letters, plurals, suffixes, etc.). Word standardization facilitates the matching process. While phrase matching might be a simple task for a person, it is not necessarily so for a computer. A person would have no difficulty in deciding that the phrases

French Teacher
Teaching French

implied the same occupation. Yet a computer program that matches words by comparing the sequences of letters within the words would have to reach a different conclusion even if the order of the words was ignored. However, it is apparent that if the suffixes "er" and "ing" were removed from the root word "teach", both phrases would appear the same (again, ignoring word order).

Such observations caused us to make a detailed study of the words in the coding manual. Frequency counts were computed for the last letter of each word in the files, the last pair of letters, the last triple and last four letters. These counts augmented by general linguistic knowledge and judgment provided the basis for selection of the "coded" suffixes. Some 49 suffixes involving one to six characters have been assigned numeric codes for identification (see Attachment 1). The following rules transform a word to its "standard" representation.

1) Change all words which appear to be plural forms (those ending in "ies", "es", or "s") to singular.

2) Remove longest suffix and save numeric suffix identifier.

3) Replace all adjacent double letters with a single letter.

4) Terminate word standardization if the result would be less than three characters.

Examples of word transformation to standard representation:

Original Word	Std. Form	Suffix Code
ELECTRICAL	ELECTR	21 (ICAL)
SUPPLIES	SUPL	42 (Y)
MEDICAL	MED	21 (ICAL)
OFFICE	OFIC	3 (E)
OFFICER	OFIC	32 (ER)

It is important to remember that the representation of a word is incomplete unless both the alphabetic part and the suffix code are present. A zero suffix code indicates that the word does not have one of the set of 49 suffixes.

A table of acceptable suffix substitutions has been developed so that alternate forms of words are recognized. For example, a "y" and "er" may be alternates, as in geography-geographer and a "y" and "ist" may be alternates for psychiatry-psychiatrist (see Attachment 1).

The coding system accesses a small list of synonyms and abbreviations. The list is primarily made up of directives from the manuals. For example, the occupation manual contains:

FOREMAN-SEE SUPERVISOR
SHELLACKER-SEE PAINTER

In such cases, Foreman is equated to Supervisor, Shellacker is equated to Painter. Other items in the list are common abbreviations. These are equated to the full words as they appear in the manuals.

Step 3. Look up in the "lexicon" all the respondent's words. The "lexicon" is a dictionary of all individual words, in standard representation, found in the data base (approximately 7200 words). Associated with each word are its I&O heuristic weights (defined in step 6), and a pointer to the data block containing all phrases with this word. Trivial words such as "A, And, For," are eliminated. Words not found in the lexicon are processed through a spelling corrector before retrying the lexicon look up. Words still not found in the lexicon are elimiated (Unrecognized words are evaluated if case is uncodeable).

Step 4. Check for response words that imply an economic sector such as "manufacturing, wholesale", etc. If such a word is found and it is consistent with the economic sector response on the questionnaire, the case is marked for a bonus score if this economic sector is found during matching. If such a word is found but it does not agree with the economic sector response, both given and implied sectors are searched for possible candidates among the phrases of the DB.

Step 5. Eliminate duplicate respondent replies to related questions.

Step 6. Choose the "best" response word with which to start the matching. The "best" word is defined as the word with highest heuristic weight.

The heuristic weight used is an adaptation of that developed by Rodger Knaus (Knaus 1981), which measures how specific a word is to a particular code. The computational procedure is the same as that used by Knaus, but the source of the word/code sets differs. He used a sample of response/code sets to develop his measure of word usefulness; we compute the values utilizing word/code sets from the coding manuals.

The motivation for adapting this technique was to develop a measure of the usefulness of the respondent's words for searching the data base. If for example, "Supervisor of Bouquets" is a response to an occupation question, "Bouquets" should be selected as the most useful word because it occurs infrequently and is associated with only one code in DB. It should have a higher heuristic than "Supervisor" which occurs many times in DB and is associated with many codes; such words by themselves, are not very useful for matching and receive a low heuristic weight. The industry and occupation weights for each word in the lexicon were developed during the data base creation phase. Experience has shown this to be a very effective tool.

The computation for the heuristic weight is

$$H = \frac{E_u - E_w}{E_w - \varepsilon}$$

where:

$$E_w = - \sum_i^n (p_i \log_e p_i)$$

p_i = proportion of occurrence of this word with the ith code

n = number of codes in which this word occurs

note: $\sum_i^n p_i = 1$

$$E_u = - \sum_i^n \frac{1}{n} \log_e \frac{1}{n}$$

= a uniform distribution (a single occurrence in each of the n industries or n occupations)

ε = epsilon, a small arbitrary positive value used to prevent division by zero

$$= \frac{n}{n+1} \log_e \frac{n}{n+1}$$

Phase 2 (Searching and Scoring)

Step 1. In the case of industry coding, if the economic sector is not retail or if the class of worker is not government, a search is made in the Company Name List (CNL) for the company's name, within the respondent's geographic area. If not found, the standard algorithms given below are used. If found and its industry code is unique, then that code is accepted unconditionally. If the company has more than one establishment in the given geographic area with differing industry codes, then the code of the establishment with the largest work force is accepted. (This rule is under review). The original file has the company names for all firms within a certain economic sector, above some threshold for number of employees and business volume in dollars. Associated with each name are industry codes and geographic locations. This file has now been examined from several viewpoints in order to develop a procedure for understanding its content and improving search efficiency. Because its present contribution seems to fall short of the needs of the system, consideration is now being given to either increasing the file by reducing the thresholds on number of employees and business volume or eliminating this file completely.

Step 2. Logical Analysis
Experience has shown that it is advantageous to review respondents' replies for certain keywords which indicate a high-frequency-of-occurrence-industry or an industry requiring special considerations. At present the full keyword list contains the following:

College
Education
Farm
Home
Hospital
Plumbing
Restaurant
School
Station (as in Gas Station)
Warehouse

When one of these words is encountered, a special logic module ("If-then-else") is utilized to analyze the full reply to see if it is possible to bypass the DB searching and scoring process and assign a classification code directly. It is possible for some entries, such as teachers, to assign both industry and occupation codes.

If classification is unsuccessful or incomplete, then DB matching is attempted. Logical analysis, when successful, is much faster than DB searching, and in cases were subtle differences exist between codes it is more accurate. But since it looks at all of the respondent's words it must be used judiciously so that the net effect is not negative.

Step 3. The DB match begins by examining all DB phrases containing the "best" response word. This examination tests ancillary information from the responses and the DB, and uses the results as filters for bypassing inappropriate DB phrases. The filters are:

- class of worker restriction. If a DB phrase is restricted to a government employee or an owner of the establishment, or some other class of worker, the respondent must meet this criterion or the record is not considered.

- economic sector restrictions. Economic sector (WHSL, RETAIL, MFG, etc.) of the DB record must match that of the respondent (given or implied).

- industry restrictions. Many occupations are restricted to a specific industry or groups of industries.

Step 4. If a DB record is not excluded by the filters, it is scored in the following manner: Each word of the candidate DB record is matched against each word in the appropriate response field (responses to a single question). Additionally, a composite or pseudo-response, representing a combination of words from different fields, is constructed exclusively from matched words and also scored. When a match occurs in a field or pseudo field the count of words matched and the heuristic weight of the word are recorded with that response field. The ultimate score is given as

$$S=M^3*(\sum_{m=1}^{M} H_m)*100/(A_r*A_d)+BONUS$$

where:

M = number of words matched

$\sum_{m=1}^{M} H_m$ = sum of heuristic weights for matched words

A_r = number of active words in respondent's phrase (total words excluding trivial words and punctuation)

A_d = number of active words in DB phrase

An exact match is best; therefore if $M = A_r = A_d$, the score computed above is doubled. There are other situations where the score gets a bonus or penalty, such as when an economic sector is reinforced or the record contains the same word more than once. However, such bonuses and penalties are not yet fully developed in practice.

If the DB record's score is above an arbitrarily designated threshold, it is considered a "candidate"; otherwise the record is rejected.

Phase 3 (Selection of Code)

At the end of a pass for a single respondent's word, the DB candidate records are arranged in order by descending scores. If the best candidate has a score above the threshold and is at least a fixed percent better than the next best candidate, it is accepted as the winner. Otherwise, the next "best" response word is selected, and the Phase 2 process is repeated.

If all the respondent's active words are used and there is no winning DB record, the original (unstandardized) responses are printed out verbatim for classification by a coding clerk.

Example of Automated I&O Coding:

The industry section below demonstrates logical analysis and the occupation section uses pattern matching and scoring.

- Questions and Responses:

INDUSTRY
For Whom did This Person Work:
PRIVATE FAMILY
What Kind of Business or Industry Was It:
PRIVATE HOME
Is It Mainly (MFG, WHSL, RETAIL, OTHER):
OTHER
OCCUPATION
What Kind of Work Was This Person Doing:
BABYSITTER
What Were This Person's Most Important Duties:
CARE OF CHILDREN
Was This Person an Employee for Private Company, Government employee, self-employed or Working Without Pay:
PRIVATE

- Develop "standard" representation for each word (not categorical responses) and look up in the lexicon for heuristic weight

			IND	OCC
ORIGINAL	STD	SUFFIX	H.W.	H.W.
Private	PRIV	13	1	2
Family	FAMIL	42	3	3
Private	PRIV	13	1	2
Home	HOM	3	1	2
Babysitter	BABYSIT	32	3	15
Care	CAR	3	1	1
Of				
Children	CHILDR	49	2	4

- The economic sector specified by the respondent is "Other". There are no words that imply a different economic sector.

- There are no duplicate replies to eliminate.

- "Family" in industry and "Babysitter" in occupation are selected as starting words for data base matching because they have the highest heuristic weights within their respective sections.

- Class of worker is "Private". Only DB records valid for private industry will be reviewed.

Industry Code:

- Logical analysis of the respondent's reply yields industry classification "761 Private Households".

1. The word "HOM" was recognized as a word which indicates a high frequency of occurrence industry.

2. The word "PRIV" was recognized and none of the following words frequently associated with "Hom" are present.

   Aged
   Convalesc
   Homel
   Med
   Remodel
   Old
   Tour
   Build
   Elderl
   Improvem
   Nurs
   Retard

Occupation Code:

- All records containing the word "BABYSIT" are selected from the database. There are two -
  "Babysitter, industry 761, code = 406"
  "Babysitter, except industry 761, code = 468"

  The second record is not a candidate for comparison to the respondent's reply because it is not for industry 761.

- The remaining record(s) is scored

  Active Words in Respondent's phrase (A_r) = 1
  Active Words in DB phrase (A_d) = 1
  Number of matching words (M) = 1
  Sum of the heuristic weights of matched words

  $$\sum_{m=1}^{M} H_m = 15$$

  $$\text{Score} = M^3 * \sum H_m * 100/(A_r * A_d) =$$

  $$= 1 * 15 * 100/(1 * 1) = 1500$$

  Since we have an exact match, score is doubled to 3000.

- The score 3000 is above our threshold. No other record for "Babysit" is a candidate, therefore the occupation code is "406 Child Care Workers, Private Household".

  Note that if the score had been below our threshold, the next set of phrases for evaluation would be those containing "CHILDR" because it is the occupation response word with the next highest heuristic weight.

4. DATA ENTRY PROBLEM AND SOLUTIONS

The basic data entry problem is how to get the responses, handwritten by the interviewer or respondent, into the computer in an efficient manner. Small to medium sized surveys are no significant problem. Most of their data items (but not typically I&O responses) are currently keyed on data entry equipment and there is sufficient time and equipment slack to allow for the keying of verbatim descriptions. Censuses and large surveys could not utilize HIOCS until a textual input mechanism is developed which would not slow the overall processing system, as would data keying of questionnaires. Ideally, the most efficient system would be one which electronically converts the handwritten text directly into computer characters. The Bureau however, is unaware of any existing system sophisticated enough to perform this task with the accuracy required.

There are currently four data entry mechanisms in various stages of analysis.

- Optical Mark Reader (OMR) with video scan (later possibly with OCR capabilities)
- Optical Character Reader (OCR) with video scan
- Optical Character Reader (OCR) using transcription
- Computer Assisted Telephone Interviewing (CATI)

The Bureau's OMR equipment FOSDIC (Film Optical Sensing Device for Input to Computers), unlike other OMR equipment, does not scan the source documents themselves. Rather, the source documents are photographed onto 16 mm microfilm and FOSDIC "reads" the microfilm images (SORIN 1982). The locations of filled-in response circles are transformed into predetermined codes and FOSDIC writes the result to a computer readable, magnetic storage medium. A current developmental program is first attempting to expand FOSDIC'S instructional repertoire to include a programmable area video scan. This would enable FOSDIC to encode a storage medium with the information necessary to reproduce the textual response as it appeared on the questionnaire, on a video screen (Hellerman 1978). A subsequent developmental effort is expected to incorporate optical textual reading. When the appropriate system is developed, it would be possible to bypass the clerical coding stage of processing and immediately microfilm and FOSDIC the questionnaires. Then, while the remaining data are undergoing other processing steps, a file containing only identification information and the encoded image of the industry and occupation descriptions, for those cases which could not be read optically, would be displayed on a series of video display units for data keyers to enter the verbatim text onto the record. After this file has been passed through the automated coding system, it would be merged with the main data file and processing would continue. This enhanced FOSDIC system coupled with a data entry system would minimize the amount of information keyed.

An alternative technology to FOSDIC, Optical Character Recognition (OCR) is also being explored. Two OCR approaches to the problem of handwritten text are under consideration. One is the transcription of the respondent's reply by trained transcribers and the other uses the video scan technique described above.

Another system which would facilitate the use of the automated coding system involves Computer Assisted Telephone Interviewing (CATI). CATI is being tested by the Bureau on several smaller-scale surveys and consists of a centralized telephone interviewing operation in which the interviewers ask survey questions which appear on the screen of the computer terminal at their work station. Responses, either individual codes or alphabetic description, are entered (keyed) through the terminal by the interviewer and are stored immediately for later transmission to a larger computer for processing. Hence, data entry under the CATI system is much like the data keying currently used for the majority of the Bureau's surveys, but since the keyer is also the interviewer, several data collection and processing activities are bypassed or combined.

5. POTENTIAL FOR FUTURE USE

When analyzing HIOCS's operational future we have to differentiate between censuses (POPULATION) and surveys. With both the key problems are cost effectiveness, timeliness and accuracy. Cost effectiveness is not a topic for discussion here. We do believe that if we can surmount the other two problems, the technology exists today to implement HIOCS in a cost-effective manner (Hellerman 1979). The level of accuracy achieved by clerical coding, and the importance of timeliness varies between census and survey. The Population Census and a few surveys have data release dates set by law or regulation. Therefore, personnel associated with these operations are very concerned about possible delays introduced by the data capture method. Censuses train a large temporary staff for classifying I&O descriptors. Surveys use a small permanent staff of experienced personnel. The I&O coding error rates reflect this difference (Scopp 1982). The different levels of acceptable data capture speed and coding error rates means that there are different points where HIOCS is a viable alternative to clerical coding.

SURVEYS:

The Bureau survey for which the accuracy of the industry and occupation data is most critical and, because of its sample size and frequency of collection, for which the total coding costs are the highest, is the monthly Current Population Survey (CPS). This survey is the source of the official government estimates of employment, unemployment, and other labor force characteristics, including occupation and industry data. Each month, industry and occupation codes are assigned to the records of some 70,000 persons. Ironically, the size of the survey and the need for extremely rapid processing make the CPS one of the least amenable of all the Bureau's household surveys to an automated coding system. This is because the data from the CPS questionnaires are converted to computer files through the use of FOSDIC. In the current CPS processing system, the reported handwritten industry and occupation entries are converted to numeric codes by clerks who then mark response circles corresponding to these numbers. These circles are, in turn, read by FOSDIC and the code transferred to the data file. Since an automated coding system requires the input of the alphabetic I&O descriptions itself, the CPS could not use such a system until it could be demonstrated that input keying would not slow the overall processing system.

Two of the four new methodologies discussed in Section 4 are now being examined to see if they may provide the input mechanism needed to allow the CPS to utilize the automated coding system. The first would use FOSDIC with the video scan enhancement. The question that must be answered is whether the time required for the keying and computer classification of I&O descriptors will be less than that currently required for clerical coding.

The other system which would facilitate the use of the automated coding system by CPS involves Computer Assisted Telephone Interviewing (CATI). Use of CATI for CPS would entail the centralization of telephone interviewing which is currently handled by the individual interviewers (approximately 1,500 work on CPS each month). Use of centralized telephone interviewing raises several problems of population coverage, survey design, and methodology, and obviously is applicable only to CPS households which are normally interviewed by telephone (about 60 or 65 percent of the sample, since personal interviews are required for certain portions of the CPS sample). If these problems can be resolved, data from personal interviews could be entered directly into the processing system via the CATI terminals and the input would be in a format which the automated coding could handle.

Although further research on both CATI and extended FOSDIC is required before the CPS could use automated coding, the remaining demographic surveys which the Bureau conducts have their data keyed (except typically I&O responses) and could avail themselves of such an automated system as soon as it is shown to be reliable and cost effective.

CENSUSES:

In terms of sheer volume, the most useful application of automated coding is the Population Census. In 1980 we clerically coded approximately 17,000,000 documents. This task was both time-consuming and expensive. Since one of the major goals for 1990 is improved data availability, the Census Bureau is examining ways to automate certain clerical processes. I&O coding is a prime candidate.

Three of the four methods of data capture discussed in Section 4 are under consideration. They are FOSDIC and OCR, both with data scan capability, and direct entry with OCR after transcription of handwritten text. The major problems with these methods are transcription or keying errors and the continued need for a large clerical staff. On the other hand, the staff would require only a generalized skill, keying or transcription, rather than the specialized skill of coding. Generalized skills lend themselves more readily to decentralization than do specialized ones. This would allow editing and coding at the point of entry and some errors could be corrected before they complicate later processing. If decentralization proves infeasible, both methods could be centralized with computer-assisted coding at the processing site.

6. STATUS of HIOCS

HIOCS is currently running in a research and development mode, utilizing the Consumer Expenditure Survey (CE) in an attempt to assess and improve its capabilities in an operational setting. Each month the Bureau's Regional Offices, as part of their normal processing, key CE industry and occupation natural language responses and transmit them through headquarters' computers to the Census Bureau's Jeffersonville, Ind. processing office for clerical coding. Once classified, the codes are keyed and transmitted to Washington for merging with the rest of the respondent's data.

The verbatim text sent for clerical coding is also classified by HIOCS. When the results of the clerical coding are received they are compared to the HIOCS codes and a listing of the differences is produced. Headquarters I&O coding experts adjudicate the differences, identifying both preferred and acceptable codes. Staff personnel then make the alterations indicated. The table below shows the latest results. These are for the second panel (only one set of improvements) of what is to be a twelve panel cycle of coding, evaluation and inhancement. There was significent improvement between the first two panels, although measurment error in the first panel makes quantifying the changes impossible.

Automated Industry and Occupation Coding
Panel November 1982

	Industry		Occupation	
	Number	Percent	Number	Percent
Cases	930	100.0	930	100.0
Coded	821	88.3	817	87.8
Acceptable	669	81.5	585	71.6
Unacceptable	152	18.5	232	28.4
Uncoded	109	11.7	113	12.2

The research and testing described above is applicable to surveys. For high volume operations such as a census, additional questions must be evaluated.

- For surveys, the interviewer completes the questionnaire: in a census it is the respondent. Does this difference significantly affect HIOCS performance?

- Automated coding assumes a three-path data capture system, not the present single-path. The paths are:

  1. OMR, OCR captured data
  2. Unreadable data, keyed and computer coded
  3. Unreadable data, keyed and manually coded

  What are the significant affects of this added complexity upon the data processing system.

- What level of computer resources, in the form of hardware, storage and time, will automated coding require?

- What are the personnel requirements to monitor and maintain the system?

To evaluate these constraints, extensive testing of data capture and coding on a large scale will be necessary. The first full-scale pretest for the 1990 census will probably be in April 1985. The Bureau hopes to start testing some kind of automated coding at that time, probably a centralized computerized coding system based on HIOCS. Other approaches, including decentralized systems, will also be tested in later pretests. In the meantime, the Bureau will began experimenting with computerized coding of samples of 1980 census questionnaires.

ACKNOWLEDGEMENTS

The authors are indebted to many associates who provided knowledge and assistance with various sections of this paper. In particular we would like to express our appreciation to Robert T. O'Reagan for his valuable and extensive work on the history of automated coding, to Richard L. Pauly and Robert F. Clark for their assistance with the section on data capture and to Thomas C. Walsh, B. Gregroy Russell, and Thomas S. Scopp for their help with the sections dealing with the potential uses of automated coding. Besides those mentioned above, several colleagues, especially Lawrence H. Cox and James L. O'Brien, read all drafts and made valuable suggestions on improving the paper.

REFERENCES

Corbett, J.P. (1972), "Encoding from Free Word Descriptions", draft memo, Bureau of the Census.

O'Reagan, R.T. (1972), "Computer Assigned Codes from Verbal Responses", Communications of the ACM, 15, 455-459.

Hellerman, E. (1978), "Large Scale Data Entry Systems", draft memo, Bureau of the Census.

Hellerman, E. (1978), "Lexicon Word Nonrecognition", draft memo, Bureau of the Census.

Knaus, R. (1981), "Semantic Decision Making," in Emperical Semantics (VOL 1) in Quantitative Linguistics (VOL 12), ed. Burghard Rieger: W. Germany: Studienverlag, Bochiam.

Hellerman, E. (1982), "Overview of the Hellerman I&O Coding System," draft memo, Bureau of the Census.

Scopp, T.S. (1982), "1980 I&O Coding Performance", draft memo, Bureau of the Census.

Sorin, M. (1982), "Data Entry Without Keypunching," "Data Entry at the Census Bureau", Lexington, Mass: Lexington Books, 4, 85-128.

Attachment 1: SUFFIX CODES

CODE	SUFFIX	CODE	SUFFIX
1	IC	25	MEN
2	ED	26	WOMAN
3	E	27	WOMEN
4	EE	28	ION
5	AGE	29	SION
6	BLE	30	TION
7	ABKE	31	PERSON
8	IBLE	32	ER
9	CLE	33	OR
10	WARE	34	IER
11	WEAR	35	ESS
12	HOUSE	36	OUS
13	ATE	37	ET
14	ITE	38	ANT
15	IVE	39	ENT
16	ING	40	IST
17	EK	41	YST
18	WORK	42	Y
19	AL	43	LADY
20	IAL	44	BOY
21	ICAL	45	GIRL
22	EL	46	ANCE
23	IAN	47	ENCE
24	MAN	48	ETTE
		49	EN

Acceptable Substitutions:

IC, ICAL
E, AL
ATE, ITE
WARE,WEAR
ANT, ENT, ET
ION, SION, TION, IVE
IAL, ICAL, Y
BLE, ABLE, IBLE
ER, OR, IER, ESS, IST, YST, IAN, Y
MAN, MEN, WOMAN, WOMEN, PERSON, LADY, BOY, GIRL