We may conclude that trigram coding can be a valuable tool for coding verbal responses. However, more research may be necessary to see how this method works in other situations.

REFERENCE

[1] Lina, M. Blaise 2.5 / Interactive Coding. Netherlands Central Bureau of Statistics, Voorburg, The Netherlands, 1993.

THE 1991 CANADIAN CENSUS OF POPULATION EXPERIENCE WITH AUTOMATED CODING

By Jocelyn Y. Tourigny and Joanne Moloney, Statistics Canada

ABSTRACT

Automation can improve the quality of coding and save resources. The paper details the 1991 Canadian Census of Population experience with a coding software developed by Statistics Canada (ACTR). The census automated coding is justified and the benefits are explored.

Keywords: automated coding; coding software; computer assisted manual coding; parsing; matching; census.

1. INTRODUCTION

The 1991 Canadian Census of Population completed the automated coding of 10 questions with write-in responses using a software called ACTR (Automated Coding by Text Recognition). In this paper we discuss the problems of coding in a census environment and the advantages of an automated coding system. We review the development of the Census coding application and ACTR.

2. JUSTIFICATION OF AUTOMATED CODING

In the context of a survey, questions requiring written responses are useful when the studied characteristic has a large set of possible response categories or when some of the outcomes cannot be predicted. Written responses allow the survey taker:

- to simplify the formulation of the question by offering the respondent fewer multiple choice questions;
- to be more objective by reducing or eliminating the artificial structure of the multiple choices proposed (and the order of the choices) thereby countering the respondents' tendency to check only the first relevant choice;
- to obtain a variety of responses that can lead to a re-examination of the classification structure and, when necessary, its modification; and
- to simplify the respondents' task because their responses are in the same medium as the question.

In order to facilitate statistical tabulations and analysis, it is necessary to group the written responses semantically using a structured classification system. This operation is called coding.

Traditionally, coding is a manual operation. Using the written response (and possibly other information provided by the respondent), and coding instructions, a coder searches for the response or an approximate alternative in the corresponding classification manual or reference material. The associated code is entered on the questionnaire. This code is then captured and used for subsequent tabulation and analysis.

Organizing manual coding of census results always rises many problems related to the specific requirements for personnel, difficulties to ensure quality and timeliness, and to integrate the coding operation into census process.

Because of that, alternatives to manual coding were sought. Automated coding was selected because of its potential to reduce the dependency on coding staff and reduce overall cost to some extent. Improvements in the quality of results arise from the predictability and consistency of computer systems.

Statistics Canada has developed an automated coding system that can meet the needs of various surveys. This generalized system, known as ACTR, is used in several surveys, the largest of which is the 1991 Census of Population.

3. AUTOMATED CODING METHODOLOGY (ACTR VERSION 1.06)

3.1 General

The methods used by the ACTR system are based in part on methods that were originally developed at the US Bureau of the Census [4] and in part on the experience of Statistics Canada in developing matching algorithms and systems for administrative files processing. The response to be coded is compared to a series of pre-coded responses, called a reference file. If a match is detected the corresponding code is recorded and the operation is complete. If not, the search continues, and an algorithm is introduced to locate the most comparable response. Once this operation is completed, the system attributes the corresponding code.

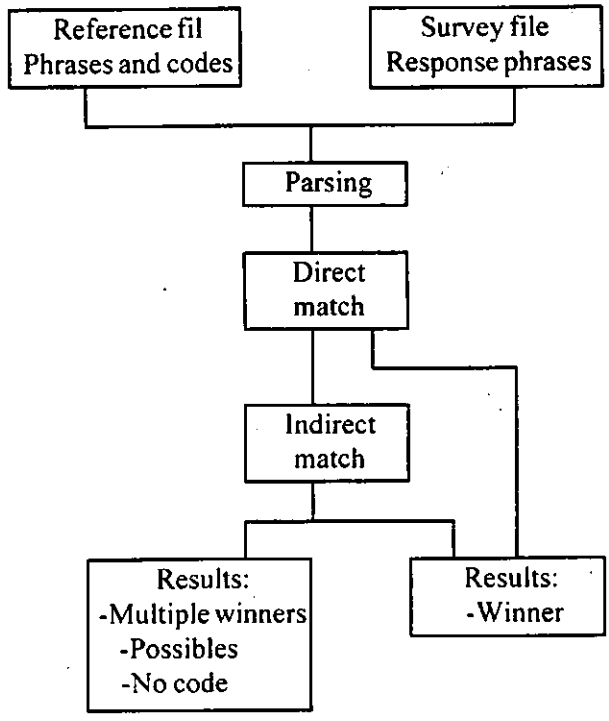
This search is made more complex because of the fact that the human language has several ways to express the same notion. Words are not always in the right order, an important word may be missing, an extraneous word may be present, a word may be a synonym or abbreviation of an expression, or the rules of punctuation and syntax may not have been respected. ACTR addresses these problems through prior processing (called parsing) of responses as well as through its two matching techniques.

Figure 1 depicts the various modules of the ACTR system that we shall describe.

3.2 Reference file

For each question to be coded, it is necessary to create a reference file consisting of typical written responses (called **phrases**) for that question and the associated numeric code. Ideally the phrases chosen are representative of the phrases most frequently observed in a matching operation. It is recommended that the phrases be retained in their original form, with errors in spelling, grammar and syntax. This file of phrases and numeric codes is integrated into a data base serving to facilitate matching operations. The reference file is constructed using entries from standard classification manuals, phrases coded by experts from a similar survey conducted previously, or a combination of these two sources as in the case of the 1991 Census of Population.

Figure 1. ACTR system



3.3 Parsing

The phrases in the reference file and those to be coded are converted in standardized form, or "parsed," in order to enable the computer to recognize, as identical, responses that are semantically equivalent. ACTR provides the user with a highly flexible parsing module. First, the phrase is considered as a continuous string of characters; it is not recognized as containing words, spaces and punctuation marks. This string of characters is analysed by the system to identify separate words. The separate words are then scrutinized and parsed; the latter stage reduces the problem of synonyms, double words, trivial words, different suffixes, etc. Annex 1 provides a list of the parsing functions offered by ACTR.

3.4 Direct matching

The parsed words of the response phrase are put in alphabetical order and the phrase is condensed to a length which averages 35% of the initial length of the phrase; the result is called the CPK (Compressed Phrase Key). This key is constructed by eliminating spaces between the parsed words and by converting individual

characters (letters and numerals) and frequent combinations of characters to bit code representation. The key is then used to search for an "exact" match in the reference file, where each phrase already has its own key.

3.5 Indirect matching

This method consists of searching in the reference file for the closest match to the response phrase when a direct match cannot be found. All phrases that have one or more parsed words in common with the response phrase are extracted from the reference file. The system evaluates each of these phrases and assigns them a "score." This score, combined with certain pre-established parameters, is used to determine whether there is a "winner" match, "multiple winners" or "possible" matches in the reference file. This method is inspired by the works of Hellerman [4] and Knaus[5].

3.5.1 Calculation of a weight for each parsed word in the reference file

The system calculates a weight for each parsed word contained in the reference file. This weight gives an indication of the power of discrimination of the word -- that is, it indicates whether the word can lead to a single numeric code.

The heuristic weight of the word is constructed in such a way that it decreases as the number of codes with which it is associated increases. The weight H of a word has the following form:

$$H = \frac{E_U - E_M + \epsilon}{E_M + \epsilon}$$

where:

$$E_M = -\sum_{i=1}^{n}(p_i * \log_2 p_i) \wedge E_U = -\sum_{i=1}^{k}\frac{1}{k} * \log_2(\frac{1}{k})$$

E_M is the entropy of the word. Entropy is a measure of the uniformity of a distribution. When a word is specific to a single code, the entropy is nil; it reaches its maximum when the word is associated with all items (that is the *n* codes) in the classification system.

p_i is the proportion of occurrences of the word in the files for the i^{th} code; this quantity is therefore a measure of the probability that given the word, the appropriate code is code *i*.

$$k = \sum_{i=1}^{n} x_i \;,\; p_i = \frac{x_i}{k} \;,\; \sum_{i=1}^{n} p_i = 1$$

x_i is the number of occurrences of the word in question in the phrases that have code i.

ϵ is an arbitrary small constant to avoid division by zero in the event that $E_M = 0$ (which corresponds to the situation where a word is specific to a single code).

$$\epsilon = \frac{k}{k+1}\log_2\frac{k}{k+1}$$

3.5.2 Calculating a score for each matched phrase

Each reference file phrase that contains at least one parsed word in common with the response phrase is considered a potential match. A scoring method was developed in order to determine the closest phrase; this score is based on the number of words contained in the response phrase that are "valid" (ie. present) in the reference file, the number of words in the reference file phrase, and the weight of the words common to the two phrases. The formula used is as follows:

$$P = \frac{(\text{number of words in common})^3 * (\Sigma \text{ weights of words in common})}{(\text{number of valid words in the response phrase}) * (\text{number of words in reference file phrase})}$$

When a response phrase matches exactly a phrase from the reference file, the formula becomes:

$$P = (\text{number of words in common}) * (\Sigma \text{ weights of words in common})$$

3.5.3 Evaluation of matches and selection of a winner

To resolve indirect matches, the user assigns values to the following three parameters:

1. MIN: lower limit of score
2. MAX: upper limit of score
3. PCNT: percentage difference

Let us assume that there are m possible matches in the reference file. The scores obtained by these phrases are arranged in descending order:

$$P_1 > P_2 > \ldots > P_m$$

As a result, four situations may arise:

(i) If $P_1 >= MAX$ and $(P_1 - P_2) / P_1 >= PCNT$

then the phrase that obtained score P_1 is the **winner** and its numeric code is assigned to the response phrase.

(ii) If $P_1 >= MAX$ and $(P_1 - P_2) / P_1 < PCNT$

then all phrases i such that $P_i >= MAX$ are considered as being **multiple winners**.

(iii) If $MIN <= P_1 < MAX$

then all phrases i such that $MIN <= P_i < MAX$ are considered as **possible** matches.

(iv) If $P_1 < MIN$

then no match qualifies.

All response phrases in situations (ii), (iii) or (iv), as well as those with no potential match in the reference file must be coded manually. During the tests prior to production, all such response phrases available are studied in order to improve the reference file, the parsing rules and the matching evaluation parameters.

3.6 ACTR performance

Owing to its use of the compressed phrase key, the direct matching technique is highly efficient, even when the reference file is very large.

To make indirect matching more effective, ACTR extracts from the reference file all the phrases that contain the word in the response phrase with the highest heuristic weight H, and determines their scores. Next, the word in the response phrase with the second highest weight is identified and, using this weight, a "maximum possible" score is estimated. If this score is lower than the MIN parameter (the score for a valid match) the process is halted. Otherwise extraction of reference file phrases and calculation of their scores continue.

4. 1991 CENSUS CODING APPLICATION

4.1 General

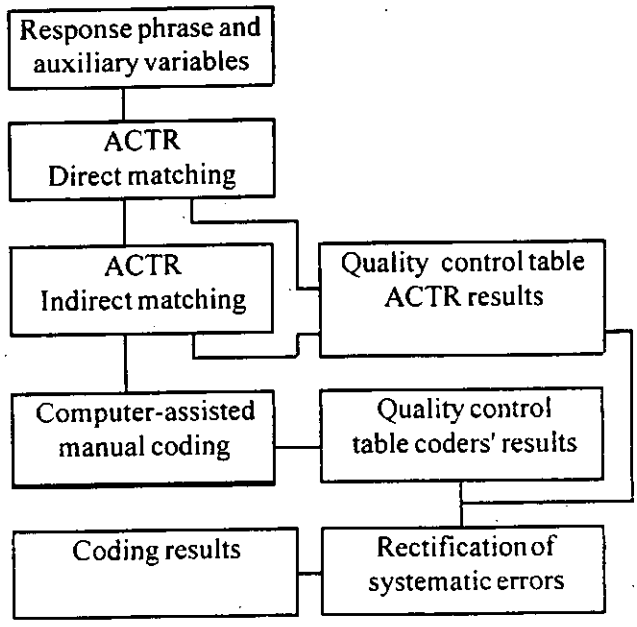
The Canadian Census of Population and Housing uses two types of self-administered questionnaires to canvass more than 10 million dwellings. When establishing the list of dwellings in his or her enumeration area, a census representative distributes a short questionnaire to 80% of the dwellings and a long questionnaire to 20% of the dwellings, following a systematic sampling. The respondent returns the completed questionnaire by mail.

The long questionnaire serves to collect information on the characteristics of individuals. The short questionnaire is an abridged version of the long questionnaire; it includes only basic questions on housing and individuals (e.g., type of dwelling, owner- or tenant-occupied dwelling, relationship to Person 1, sex, date of birth, legal marital status, common-law status and first language learned). To respond to a question, the respondent must mark a circle, write a number or write in a response.

Some write-in responses are coded manually in preparation for data capture. All the information on short and long questionnaires, except for write-in responses already coded, is captured in a single operation over a four-month period. For each variable subject to automated coding, the write-in response as well as auxiliary variables relating to the person and other occupants of the dwelling are transferred to a data base to facilitate the coding operation.

The 1991 Census automated coding application is illustrated in Figure 2. The application is highly integrated. It encompasses automated coding by ACTR, computer-assisted manual coding, quality control of the two types of coding, and rectification of systematic errors. No return to the actual questionnaire is necessary.

Figure 2. Census coding application modules



The 10 questions subjected to automated coding are shown in Appendix B. Of these, 12 similar but customized applications were established.

4.2 ACTR - direct matching

Only the response phrase without its auxiliary variables is used for automated coding. The system identifies unique response phrases. It is this unique phrase that is parsed and matched with the parsed phrases in the reference file. If there is a match, all the response phrases corresponding to the unique phrase receive the same code and the result is entered in the quality control table for ACTR results.

For the Census, the automated coding of 9 of the 10 questions was done solely by way of this matching method. Only the **Place of Residence 5 years ago (write-in Canadian cities, towns and municipalities)** question also used indirect matching to increase its automated coding matching rate.

4.3 ACTR - indirect matching

All unique phrases that are not coded via direct matching are then subjected to the indirect matching method. Information concerning the "multiple winners" and "possibles" (the matched reference file phrase, the corresponding code and the score) is forwarded to the computer-assisted manual coding. If there is no match, or if there are only matches with scores below the minimum score MIN, no information from ACTR is forwarded.

4.4 ACTR - notes on execution

A number of applications shared the same reference files and the same parsing strategies. These files were constructed using entries in classification manuals, a sample of responses from the 1986 Census and responses from ongoing household surveys. The files contained both English and French entries which did not cause deterioration of results.

Since the coding application was run daily, it was possible to analyse ACTR results and uncoded phrases regularly. During a four month period, reference files were updated five times in order to increase the automated matching rate and the quality of the results. No improvement of parsing strategies was permitted because the impact on the quality of the results was unforeseeable.

4.5 Computer-assisted manual coding

For phrases failing automated code assignment, the computer searches the original file of response phrases (ordered alphabetically) and prepares batches of 200 uncoded phrases. Coders do not have access to the original questionnaire, but the following information appears on two screens (see Figures 3 and 4). On the first screen the coder sees the phrase to be coded,

Figure 3. FIRST.SCREEN

MANUAL CODING - MAJOR FIELD OF STUDY

RESPONSE PHRASE	Type	Code
RENAISSANCE ARCHITECHTURE	____	____
Phrases provided by ACTR	**Codes**	**Choice**
ARCHITECHTURE	267	
ARCHITECTURE D'ART	048	____
BOAT ARCHITECHTURE	308	____
		____

Data of each household member for the same question

Check boxes : .....
Write-in phrases : .....

PF1 = Help ..... PF9 = Referral

Figure 4. SECOND SCREEN

MANUAL CODING - MAJOR FIELD OF STUDY

Number of years

Elementary and secondary school :	12
University :	4
Other schools :	NONE
Education during the past 9 months :	NO
Diploma :	SEC / CERT BACC MASTER
Economic activity :	8531 TEACHING / UNIVERSITY
Occupation :	2711 TEACHER / UNIVERSITY
Major field of study :	RENAISSANCE ARCHITECHTURE
Relationship to person 1 :	PERSON 1
Date of birth :	30 / 01 / 1927
Sex :	M

the ACTR results (matched phrases and associated codes), and lastly the responses of other members of the household to the same question. On a second screen the coder can obtain the person's responses to other questions. The coder may either select one of the ACTR results, enter a code based on a classification manual or refer the case to an expert. Each time the coder selects a code, the system shows at the bottom of the screen the official rendering from the classification manual; the coder must read and confirm the code. The result of the coding is entered in the quality control table for the coder's results.

The computer electronically transfers the phrases referred to the expert on duty. The expert has on-screen access to additional information, such as the ACTR scores and auxiliary information for all other members of the household. In addition, he or she can consult more specialized reference manuals.

4.6 Quality control table for ACTR results

Quality control for automated coding has the same objectives as those of traditional coding. However, it differs in scope, since much more information on the operation is available and this information may easily be altered.

Every aspect of quality control exploits the systematic nature of automated coding, since a phrase always receives the same code if there is no human intervention. Thus, the examination of a single occurrence of a phrase is sufficient to determine its quality. The conclusions as to its quality extend to all replicates of this phrase.

The quality control (QC) table contains one entry for each phrase-code pair. A status indicator is associated with the pair. Its value is 1 for a pre-approved pair, 2 for a pair that has been verified and found valid, 3 for a pair that has been verified and found invalid and 4 for a not verified pair. During production, each new automatically coded phrase-code pair is added to the table, and the frequency of occurrence is increased with each repeated pair.

Since the initial entries in the reference file have been intensively tested, all pairs included in this file are entered on the QC table with pre-approved status, and they are not verified. This makes the quality control more efficient.

The other pairs are sampled on a priority basis. As soon as a phrase-code pair has a frequency of three or more, one of the replicates is selected and coded by a coding clerk.

The system compares the code assigned by ACTR with the one supplied by the coding clerk. If the codes correspond, the pair is considered valid. Otherwise the case is submitted to another coding clerk. If the new code corresponds to the ACTR code, the pair is considered valid. If it corresponds to the one assigned by the first coder, the pair is considered invalid. Finally,

if it does not correspond to either of the codes, the case is sent to a referral coder.

This type of quality control identifies the differences between the manually established code and the ACTR code, and it assists in detecting operational problems in the two types of coding.

In addition to facilitating sampling for quality control, the QC table serves to regularly calculate error rates. The subject matter specialist may also scrutinize phrase-code pairs not verified and determine the coding quality.

4.7 QC table for coders' results

The QC table for coders' results contains one entry for each response phrase processed. This phrase is accompanied by the code assigned by the coder, a batch number, the coder's identification number and the final code assigned after quality control.

The objectives of the quality control are to determine coder performance, identify problem areas, ensure that quality objectives are met, provide feedback on the process and prevent the recurrence of error.

The quality control method used is that of sampling by attributes, with 100% rectification of rejected batches. In practice, 5 phrases from a batch of 200 are verified by a coding clerk. As in the case of quality control of ACTR results, there is no further inspection if the codes correspond. Otherwise, it is necessary to bring in a second coding clerk and lastly a referral coder to determine the correct code.

A batch is rejected and recoded if only one phrase has an erroneous code.

The code that appears in the census file is either the code established during inspection or the original code if it has not been inspected. Error rates are calculated regularly.

4.8 Rectification of systematic errors

The two QC tables contain the history of the automated coding and the manual coding. When analysing these tables, the subject matter specialist identifies the errors to be corrected. The analysis may also lead to a change in the classification system to reflect a new reality. The census application includes a rectification module that is used at the end of production, immediately before the results are integrated into the main census processing data base.

The systematic error rectification module acts globally on erroneous phrase-code pairs and extends its action to all replicates of each pair. Detailed reports of the actions taken are produced in order to ensure proper control of this operation.

4.9 Results and observations

4.9.1 Coding volume and match rate

For the presentation of results, the responses to the 10 questions subject to automated coding were grouped under 7 variables which corresponded to separate reference files and parsing strategies. Table 1 presents these variables and some processing statistics relating to them.

Table 1. Automated Coding - variables and statistics

Variable	Processed	Matched by ACTR	ACTR rate	*CAC* Manually coded
Ethnic origin	1,160,491	1,062,015	91.51%	98,476
Language	5,998,021	5,741,294	95.72%	256,727
Registered Indian	236,501	169,675	71.74%	66,826
Place of residence 5 years ago (city/town/muni.)	1,042,951	793,425	76.08%	249,526
Major field of study	1,905,959	1,485,196	77.92%	420,763
Province - Country - Territory	880,077	821,510	93.35%	58,576
Religion	4,859,569	4,752,021	97.79%	107,548
Total	16,083,569	14,825,136	92.18%	1,258,433

Of the 16 million responses sent for automated coding, the ACTR system coded 14.8 million or 92.18% (the match rate). The remaining 1.2 million cases were resolved through computer-assisted manual coding.

The match rates fall into two main clusters: in the 71% to 78% range and in the 91% to 98% range. The disparity in match rates by variable may be explained by the volume processed, the response variation, the length of the responses, respondents' use of abbreviations, changes in national boundaries owing to the collapse of the Communist bloc and the fact that certain variables (for example, a municipality name associated with several codes) were deliberately sent for manual coding where auxiliary information could be used to obtain the correct code.

The **Registered Indian** question was new, and it was difficult to anticipate the responses, particularly since various names have recently undergone numerous changes. The **Place of residence five years ago** variable avoided the use of duplicate place names by not including them in the reference file. Duplicate place names include geographic locations that have the same name either within a province or, if the province is not identified, in more than one province. In addition, a name such as "Québec" was excluded since it could refer to either the province or the city. The **Major field of study** variable had a number of quite varied responses, diverse nomenclature, the use of abbreviation and wordy responses. The problem with wordy responses is that errors in any one word may prevent direct matching, the only type of matching allowed for this variable. In addition, it was not possible to list all possible spelling variations and abbreviations for these responses. Lastly, lengthy responses are more prone to keying error in the data capture operation.

4.9.2 Update of reference files

During production there were five reference file updates. It is estimated that they raised the match rate by 2 percentage points, which reduced the manual coding volume by approximately 25%. In some cases reference file phrases were deleted because they were found to generate errors.

4.9.3 Analysis of the QC table for ACTR results

As noted above, all unique phrase-code pairs have one of the following statuses: pre-approved, verified and found valid, verified and found invalid, and not verified.

The term "invalid" as used here indicates that there is a difference between the ACTR code and the code identified during quality control. Differences may be due to various factors: erroneous codes in the reference file, overly parsed phrases, or coders not having the latest instructions or making errors of judgment or oversight. Another cause of differences is that the response may be associated with several codes. Thus what we are measuring here is a gross difference that must be analysed before a rectification is initiated. The analyst also has the task of detecting any errors that have been missed during quality control.

Table 2 presents the volume of phrases by status. More than 87% of the phrases coded by ACTR were pre-approved. Fewer than 1% of the phrases were identified as having an invalid code.

4.9.4 Quality control resources

The resources allotted for quality control provided for verification of 3.0% of the ACTR-coded responses and 10.0% of manually coded responses. The final rates were 0.251% (Table 2: [2,705 + 34,499]/ 14,825,136) for automated coding and 10.02% for manual coding.

The rate of 0.251% can be attributed to the high frequency with which pre-approved phrase-code pairs occur and the fact that each unique pair was selected and verified only once. Such an inspection strategy is impossible in a traditional quality control operation. This rate thus indicates that using all the information produced by the system can increase the efficiency of the inspection without compromising quality.

Table 2. Results of quality control - all variables

Status	Unique Pairs	Frequency	Freq. (%)	Control (%) total
Pre-approved	14,787	12,898,773	87.01	
Verified and invalid	2,705	89,743	0.61	0.018
Verified and valid	34,499	1,735,931	11.71	0.233
Not verified	82,128	100,689	1.67	
Total coded by ACTR		14,825,136	100.00	

Table 3. Average frequency of phrase-code pairs by variable and by status

Variable / status	Pre-approved	Verified and invalid	Verified and valid	Not verified
Ethnic origin	528	12	27	1
Language	1,906	167	128	1
Registered Indian	103	13	37	1
Place of residence 5 years ago (cities/towns)	---	19	44	1
Major field of study	180	16	29	1
Province - Country - Territory	588	393	38	1
Religion	4,252	25	105	1
All variables	872	33	50	1

Table 3 illustrates, for each variable, the average frequency of occurrences of unique phrase-code pairs coded by ACTR.

The average frequency of pre-approved phrase-code pairs is 872. The most interesting frequency is that of pairs that were verified and found invalid, with an average of 33. This means that correction of one of these pairs rectifies an average of 33 errors.

The average frequency of pre-approved phrase-code pairs is 872. The most interesting frequency is that of pairs that were verified and found invalid, with an average of 33. This means that correction of one of these pairs rectifies an average of 33 errors.

For the next Census, the goal will be to pre-approve as many pairs as possible in order to minimize the resources allocated to quality control. The resources thus freed up can be used to better analyse the two quality control tables.

4.9.5 Rectification of systematic errors

Approximately 94,000 codes were corrected by the rectification module. The codes were obtained from the two types of coding (automated and manual). Most of the rectifications resulted in an improvement in quality. For the **Ethnic origin, Language** and **Province - Country - Territory** variables, several codes were changed to reflect the new world reality, which changed considerably between the production of the questionnaire and the end of processing of the Census data.

Our estimate of the final quality for the two types of coding is a combined error rate of less than 1%; manual coding is the main source of errors. However, the rate achieved is remarkably low, since in earlier censuses the error rate was in the range of 4% to 8%.

5. BENEFITS OF THIS NEW CODING PROCESS

In its first large scale use, automated coding successfully met its objective of reduced coding staff and costs, with improved data quality. Within 4 months of processing, the Automated Coding project was able to reduce the number of coders from 600 to 25 and save over $3.5 million on an estimated budget of $5.9 million, while achieving an outgoing error rate of less than 1.0% - a fraction of the error rate associated with manual coding. The dollar savings for coding takes into account the development of the census application systems, the development of reference files and parsing strategies for the automated portion and coding and training material for the computer-assisted portion; these savings were offset by a charge of $900,000 for the data capture of the write-ins.

Additional benefits are grouped under 4 headings; these are:

- Maximum control retained by the subject matter (SM) specialists

SM specialists developed the reference files and the parsing strategies using specialized tools provided by ACTR. During production, they regularly monitored the write-in responses that were not coded by the ACTR system and were able to add entries to the reference files to improve the match rate, and to remove or modify entries in the reference files to improve the quality of the results. The quality control processes provided additional feedback to the SM specialists leading to immediate corrective actions.

SM specialists directly trained the manual coders for CAC because their number was small and the operation was centralized. The computer-assisted portion of the training ensured consistency across coders. SM specialists were present during processing to solve difficult or unanticipated cases and to quickly update coding instructions when required. This would have been impossible if manual coding was used and the SM expertise would have been lost.

The CAC system was programmed to implement certain complex procedures such as enforcing the use of expert referral for multiple responses (e.g. "Polish French" language spoken at home).

The more controlled environment permitted the SM specialist to design and implement partial action at the coding stage that could be completed at the imputation stage. An example related to the coding of the mobility variable is described in [8]. If inadequate information in the response leads to a match with duplicate place name (ie. the write-in is associated with more than one place name), a pseudo-code is assigned. The census imputation process attributes a final code at random based on the frequencies of the correctly coded places related to the duplicates.

At the end of the process, after a review of the files produced by the quality control systems and the analysis of other system-generated files, the SM specialists were able to make global enhancements to the results. This procedure can consistently correct frequent errors or implement a modification in the classification.

- More efficient quality control and certification

The fact that the automated coding system repeats its coding actions can be exploited. Only one replicate of a unique write-in phrase/code combination need be subject to quality control verification. Moreover, combinations of write-in/code found in the reference file before the production start are exempted from quality control (pre-approved) because they have been reviewed and certified accurate by SM specialists. These two measures allow savings that can be used to verify less frequent combinations.

For the CAC operation, the quality control plan is completely automated; the computer selects the sample and sends it electronically to another coder for verification. For each write-in in the sample, the computer compares the codes and determines if the original code is correct. Depending upon the number of original codes in error, the computer determines if a work load is accepted or rejected and in the latter case sends the work load to be recoded. Feedback to the coder can be provided as needed throughout processing in a timely manner.

The computer allows a quality control sampling plan for each coder and more elaborate selection scheme (such as sampling at a higher rate codes known to be prone to error). These are possibilities for our next census.

For both the automated and the CAC systems, all the quality control results are on files that can be reviewed by SM specialists. Quality control statistics can easily be tabulated and forwarded to management, other subject matter specialists and coders. Quality problems identified can be corrected globally.

At the end of the process, when reviewing the code distribution against another source of statistics, if the frequency of a code assignment is "suspect", it is easy to review all the write-in entries associated with it; the entries can be printed and analysed, diminishing the need to return to the original questionnaires.

- Better management of the coding process

Because the information is in electronic format, it is easier to predict the volume of work at each step. Regular monitoring reports can be generated from the files produced by the systems. Ad hoc analyses and reports can be easily generated.

- Potential for improvement in the next application

All the write-ins and their corresponding codes are available for other uses - other surveys or future censuses. Frequently occurring write-ins can be added to the reference file and/or the coding manual in order to ensure better coverage and coding of these responses; unused reference file entries can be removed. New parsing strategies and matching techniques can be developed and refined using as test data all or a sample of the write-ins.

Edit rules could be devised to identify write-in responses that are provided in error and should therefore not be coded, responses for which a code should be imputed, as well as responses that should be forced to the CAC operation.

Realistic 'test decks' could easily be created to train the CAC coders.

Future recoding of the data when a new classification system is introduced is a possibility. To answer a specific request, it is feasible to recode in more detail write-in responses corresponding to a certain code.

Finally, someone can exploit the capacity of the computer to track and record paths taken by the computer or the coder for determining the code. This audit trail information can be useful for streamlining complex coding applications.

6. CONCLUSION

The use of automated coding for the 1991 Census was an outright success on which to capitalize for the 1996 Census.

Our intentions for the 1996 Census are as follows:

The ACTR software will again be used, but it will undergo certain changes to increase its versatility. It will have the option to specify the order of functions in the response parsing process, to retain the original word order when creating the compressed phrase key used in direct matching, and to retain duplicate words in parsing.

The 1991 coding applications will be moderately enhanced to make them more effective. The reference files and parsing strategies will be updated. A new module to be situated at the beginning of the application is being considered; it will decide whether a response should be subjected to automated coding, or be assigned a provisional code indicating that there is insufficient information to code it. Lastly a classification manual will be available on screen so as to facilitate manual coding.

Two new questions will be coded: Relationship to Person 1 and Place of Work (coded at the block level). For these questions, the coding application will be more complex and will utilize ACTR and other softwares to match files or edit codes (see [9]).

The challenge for the 2001 Census will be to provide automated coding for the last two questions that have write-in responses: Industry and Occupation. Ironically, the original intention when ACTR was developed was to code these two questions.

ANNEX 1

PARSING OF PHRASES

The ACTR automated coding software contains a module that allows for the parsing of phrases from the reference file and the survey file. It offers a fixed sequence of fourteen functions which, depending on the coding application, may or may not be used. The first four functions identify the words of the phrase; the other ten functions parse these words. For each function used, the subject matter specialist must provide a list of valid characters, words, replacement words or suffixes.

Processing of text:

The phrase is treated as a continuous string of characters, so as to be able to eventually identify separate words.

Function 1: exclusion clauses - For the phrases in the reference file, the text that indicates an exclusion clause (for example, "clerk (except in the armed forces)") must be excluded, since respondents do not express themselves in this manner. The result will be identical parsed phrases in the reference file that will lead to "multiple winners" matches. ACTR will not assign a code; rather, these matches will be routed to a coder who will have to decide on the appropriate code.

Function 2: deletion strings - Serves to eliminate extraneous characters, such as apostrophes, which would be interpreted as word delimiters by function 4.

Function 3: replacement strings - Serves to replace an abbreviation by one or more words, since otherwise the meaning of the abbreviation will be destroyed by function 4. For example, "T.V." is replaced by "television."

Function 4: word delineation - If a character is not in the list of valid characters for a word, this function indicates the beginning or end of a word. For example, if only numerals, letters and the hyphen are valid, the following two phrases will be divided into two words: "T.V." = T V, "English/French" = English French; the phrase "Electrician's Apprentice" will be divided into three words.

Processing of words:

The phrase is treated as a collection of words. Consequently the following functions apply to the words considered individually.

Function 5: hyphenated words - Serves to retain as a single word two words that together have a specific meaning, such as "post-secondary." If the hyphenated word is not in the list, it is split into two words; otherwise it is replaced by a new word.

Function 6: invalid word characters - If a word is made up of a character string that makes it unintelligible, it is deleted without further consideration. In some applications, this function is used to delete words containing numeric characters.

Function 7: replacement words - This function operates in the same way as function 3. The main difference is that the search is restricted to whole words, as opposed to word parts. This function ensures that two synonymous words are recognized as being the same for matching purposes. It can also be useful for correcting common spelling errors.

Function 8: double words - If two words, when taken together in a certain order, have a particular meaning, this function serves to replace them by a single word. For example, the two words "radio" and "active" are replaced by "radioactive," and the French "garde" and "malade" are replaced by "infirmier." This function can resolve spelling inconsistencies and prevent the word order from being altered as would occur in the construction of the "compressed phrase key" in the case of a direct match.

Function 9: trivial words - an extraneous word such as an article or a pronoun does not contribute to the semantic content of the phrase and can be deleted without further consideration.

Function 10: root words - Functions 11, 12 and 13 may operate in such a way as to reduce two semantically different words to the same root. This function examines words to identify root words. If it finds one, the whole word is replaced by a substitute word, and the following three functions are not activated.

Function 11: replacement suffixes - A word is scrutinized from right to left to find the longest form of suffix listed. If such a suffix is detected, it is replaced by the approved substitute. For example, the plural marker may be eliminated so that the suffix is recognized by function 12. Thus the ending "ies" is replaced by "y".

Function 12: suffixes - Usually a suffix does not change the semantic content of a word. This function scrutinizes a word from right to left to find the longest form of suffix listed, such that once the suffix is removed, the word contains at least five characters. If a defined form of suffix is detected, it is deleted. Examples of suffixes are "able", "alist", "ian" and "er".

Function 13 - duplicate letters - the deletion of double consonants or vowels does not usually change the semantic content of the word. This deletion can eliminate spelling mistakes or data capture errors.

Function 14 - duplicate words - Only one occurrence of each parsed word is retained in the parsed phrase.

ANNEX 2

1991 Census Questions Subject to Automated Coding

First language learned

What is the language that this person **first learned** at home **in childhood** and **still understands**?

Response: if the language is other than English or French, the person specifies which one.

Note: This question appears on both the short and the long questionnaires.

Home language

What language does this person speak **most often** at home?

Response: if the language is other than English or French, the person specifies which one.

Non-official languages

What language(s), **other than English or French**, can this person speak well enough to conduct a conversation?

Response: the person may specify up to three languages.

Place of birth

Where was this person born?

Response: if born in a country other than the six countries mentioned, the person must state which country.

Ethnic origin - Ancestry

To which ethnic or cultural group(s) did this person's ancestors belong?

Response: if the person belongs to a group other than the 15 groups mentioned, he or she may specify up to two other groups.

Registered Indian

Is this person a **registered Indian** as defined by the Indian Act of Canada?

Response: if the "yes" box is marked, the person specifies the Indian band or First Nation.

Religion

What is this person's religion?

Response: the person specifies a denomination or religion or marks the "No religion" box.

Place of residence one year ago

Where did this person live **1 year ago**, that is, on June 4, 1990?

Response: if the person did not reside at an address in the same province/territory,he or she must specify the other province/territory or the name of another country.

Place of residence five years ago

Where did this person live **5 years ago**, that is, on June 4, 1986?

Response: if the person did not reside at an address in the same city or town, he or she must specify the name of the other city or town or the name of another country.

Major field of study

What was the major field of study or training of this person's **highest** degree, certificate or diploma (**excluding** secondary or high school graduation certificates)?

Response: the person indicates that the highest diploma is a secondary/ high school certificate or specifies a major field of study or training.

REFERENCES

[1] ACTR (Automated Coding by Text Recognition) Version 1.06 - User Manuals

[2] Ciok, R. The Use of Automated Coding in the 1991 Canadian Census of Population, *Paper presented* at *the 1991 Annual Meeting of the American Statistical Association*, Atlanta, Georgia, 1991.

[3] Ciok, R. The results of automated coding in the 1991 Canadian Census of Population. *Paper presented at the 1993 Annual Research Conference, organized by the US Bureau of the Census*, 1993.

[4] Hellerman, E. Overview of the Hellerman I&O Coding System. Internal document, US Bureau of the Census.

[5] Knaus, R. Pattern-based Semantic Decision Making, *in Empirical Semantics*, edited by B Rieger, Bochum, West Germany, 1981.

[6] Knaus, R. Methods and Problems in Coding Natural Language Survey Data, *Journal of Official Statistics*, Statistics Sweden, Vol 3, No. 1, 1987, pp. 45-67.

[7] Lyberg L. and Dean P. International review of approaches to automated coding. *Work session on Statistical Data Editing.* Geneva, 28-31 October 1991.

[8] Norris M.J. and Coyne S. Automated coding of Mobility Place Name data for the 1991 Census. *Symposium 91 -Spatial Issues in Statistics*, 1991, pp. 83-94.

[9] Tourigny, J., Moloney J. and Miller D. The 1991 Canadian Census of Population experience with automated coding. *Work session on statistical data editing,*, Stockholm, Sweden, 1993.

[10] Wenzowski, M.J. ACTR - A Generalized Automated Coding System. *Survey Methodology*, 14, 1988, pp. 299-307.