Coding of textual responses: various issues on automated coding and computer assisted coding

Stefania Macchia, Manuela Murgia

ISTAT – Integration and Technical Standard Department – Rome – Italy

Abstract

The Italian National Institute of Statistics adopted automated and computer assisted coding systems in order to overcome the problems connected with manual coding activity of textual responses of survey questionnaires, being this latter activity very time-consuming, costing and error prone. The chosen systems are ACTR v3 and BLAISE that are based on two different coding philosophies: the Automated Coding (AUC) the first and the Computer Assisted Coding (CAC) the second. The two philosophies have different scopes in that AUC aims at maximising the number of unique codes assigned to the verbal responses whereas CAC aims at providing the operator with as much assistance as possible. This paper compares the two systems highlighting the basic differences and indicating when a coding procedure is more suitable than the other. Besides, it describes the experiences made by ISTAT and the obtained results.

Keywords: Automated Assisted Coding; Textual Data.

1. Basic methodologies of coding systems

Coding of verbal responses could be defined as the attribution of numeric codes to statements according to a manual of official classification. Verbal responses resulting from statistical surveys are usually manually coded, but this operation is generally very time-consuming, costly and error prone, especially for large amount of data. For this reason and for simplifying as much as possible the coding operation, a lot of National Statistics Institutes decided to adopt automated coding systems based on specific software. The fundamental part of these computerised approaches is a database (*dictionary*) containing words or phrases associated with numeric codes, that represent the possible values to be assigned to the variables entering the coding process. The dictionary has to contain the definitions of official classifications – that constitute the starting point for the construction of the database itself – as well as the empirical responses coming from previous surveys or pilot studies. This mixture of official and empirical definitions helps in assigning a code in that the coding procedure can take into account both the official and the common language. Besides, a continuous update of the dictionary is necessary to cover the variability of the spoken language – a lot of different words to express the same concept – and also to take into account its continuous changes.

Several computerised instruments perform automated coding. In the late sixties, the US Census Bureau realised different coding systems, called "*dictionary algorithms*", that build the dictionary on the base of a large sample of verbal responses manually coded by experts. The simplest algorithm for automated coding software builds the dictionary searching for an exact match, that is, searching for the verbal description in the expert coded file that perfectly corresponds to the verbal response to be coded. One of these algorithms is described in O'Reagan: the computer analyses the expert file and decides whether the presence or absence of certain words (*key words*) indicates unique code numbers. Whether or not a description can be included in the dictionary depends on the value of three user-defined parameters. One of them (N) indicates the number of times a description occurs in the expert coded file (if the occurrence frequency is greater than N then the description is inserted in the dictionary) and the other two represent threshold values to avoid the presence of equivocal and incomplete descriptions. Users have to set these parameters in order to maximise to coding rate, keeping low the coding error. Corbett describes a different dictionary algorithm: a description belonging to the expert coded file is included in the dictionary if it contains a "*classifier*", that is to say, a word or a set of words corresponding to a specific code and whose occurrence is not lower than a defined level.

Automated coding is also realised by means of the so-called '*weighting algorithms',* that are a bit more complex than the previous ones. They assign to each single word of the input statement a weight that indicates how much a word is informative; the calculation of the weight is based on the occurrence frequency of each word in the dictionary. Afterwards, the computer searches for the input verbal response inside the dictionary: if no exact match is found then it analyses those descriptions that are "similar" to the input one and chooses the one with the highest weight, thus realising a '*partial match*'. This feature – *partial match*- represents the main difference between the *dictionary* and the *weighting algorithms.*

More articulated coding systems have been developed subsequently. Some of them - like BLAISE, Netherlands CBS - performs a partial match for both entire word and sub-strings, that is, for groups of consecutive letters of a word, thus widening the possibility of assigning the right code. Other more recent and sophisticated instruments use the so-called "*artificial intelligence*" to realise the automated coding. One of these is the "Connection Machine" – Thinking Machine Corp.- that is a computer working with thousands of processors in parallel (each representing a category – group of codes - of the official classification) that search for a code simultaneously. The peculiarity of the Connection Machine relays in its *memory based reasoning*: when searching for a match for a new input verbal response, the PC recalls codes that were attributed to similar past descriptions.

2. How to realise a coding strategy

Generally speaking, the coding activity can be performed according to two coding procedures, depending on two possible ways to use the computers:

1. the "automated coding" (AUC).
2. the "computer assisted coding" (CAC).

1. (AUC). The computer assigns codes to the verbal responses working in "batch" processing. As this technique could not be expected to assign a code to all the input statements, then a manual coding or an assisted coding procedure is required after this step to assign codes to the non coded responses.

2. (CAC). The operator assigns codes working interactively with the computer, that gives him a support in "navigating" inside the dictionary to search for codes to be assigned to the input descriptions. For example, once the operator wrote the verbal response on the PC-video, the machine shows him all those dictionary descriptions that could match with the input statement (only one description is shown if an exact match exists); the operator would choose among them, assigning the most suitable code. It could be said that the main characteristic of a CAC system is the combination of the human mind abilities and the

computer potentialities.

The difference between the two procedures relays on their final aim and coding approach. The final aim of AUC procedure is to maximise the number of unique codes assigned automatically to the input statements, whereas the CAC aims at providing the operator with as much assistance as possible. As a consequence, the coding approach of the two systems is different:

- AUC aims at extracting a single description from the dictionary matching with the input statement;

- CAC shows different descriptions (also slightly different from each other); it is important remembering that the operator works interactively with the PC and can navigate inside the shown descriptions, choosing the most suitable one. Besides, CAC allows the usage of other survey information to support the assignation of codes.

These differences imply that the phase of data coding, in the process of data collection, can be realised in different moments:

- AUC can be performed after the interview, that is, when data collection is over;
- CAC can also be used during the interview, following step-by-step data collection.

The decision about which is the most suitable coding approach to be adopted depends on different correlated factors, that is:

1. the survey technique:
   - computer assisted with the operator (C.A.T.I. - *Computer Assisted Telephone Interviews* -, C.A.P.I. - *Computer Assisted Personal Interviews*);
   - computer assisted without the operator (C.A.S.I – *Computer Assisted Self Interviews*);
   - traditional Paper and Pencil Technique (P.A.P.I.);

2. the amount of data to be coded:
   - a large number;
   - a small number;

3. the interview length:
   - short interview;
   - long interview;

4. the structure of the classification in conjunction with the variability of the verbal responses:
   - simple classification structure;
   - complex classification structure and high variability of verbal responses.

The structure of a classification can be represented as a tree with branches, sub-branches and leaves. Branches represent general levels of classification that are hierarchically higher than sub-branches and leaves, that represent detailed levels of classification. Therefore, for a simple classification structure is meant a tree with branches, non-or few sub-branches and no leaves, whereas for a complex structure it is meant a tree with all its components. Examples of simple and complex classification structure are respectively the "*Country Classification*" and the "*Occupation Classification".*

Simple Classification	**Complex Classification**
Country	**Occupation**
1. France	2. Intellectual, scientific and high specialised professions
2. German	2.1 Math, physics and natural sciences specialists and similar
3. Great Britain	2.1.1 Math, physics and natural sciences specialists
4. Italy	2.1.1.1 Physicians and Astronomers
5. Spain ….	2.1.1.2 Chemists …..

Table 1: Example of classifications with different levels of complexity

Combining the above-mentioned factors, it is possible to see whether one procedure is more suitable than the other. This combination can be analysed in two alternative situations, deriving from the moment of the implementation of the coding activity:

1. coding phase during data collection
2. coding phase after data collection.

2.1. Coding phase during data collection

The following table shows which is the most appropriate coding solution to adopt when computer data capturing is performed by an operator.

	Interview length	
Classification structure	**Short**	**Long**
• **Simple**	CAC	CAC
• **Complex & high responses' variability**	CAC	No data coding (coding after data collection)

Table 2: Survey technique: computer assisted with the operator (C.A.T.I., C.A.P.I.)

In general, as it can be seen, it is advisable to use CAC during the interview with the operator because:

- coded data are available for processing as soon as data collection is over;
- a higher quality of the coded data is also guaranteed by the contact with the respondent that can provide the operator with eventual further explanations about the given answer;
- the previous point implies that, during this activity, the operator will "train himself" in getting an answer with an informative content sufficient to be coded.

But, if the interview is long and the coding activity during the interview would amplify its lengthiness, then it is better not to use CAC and make the data coding at the end of data collection (this is especially true for complex classifications). In this way it would be avoided:

- a too high number of uncompleted interviews – respondents deny their cooperation to the operator;
- errors in coding, due to the operator's need to speed up the interview.

As shown in table 3, the situation is different when a computer assisted technique without operator is adopted for data capturing.

Classification structure	
• **Simple**	CAC
• **Complex & high responses' variability**	No data coding (coding after data collection)

Table 3: Survey technique: computer assisted without the operator (C.A.S.I.)

In this case, the coding activity during the interview – made by the respondent himself - strictly depends on the classification structure. It can only be used if:

- the classification structure is simple;
- the codes to be assigned belong to only one branch of the classification, that is to a high hierarchical level.

2.2. Coding phase after data collection

Data could be collected by a computer-assisted technique (C.A.T.I., C.A.P.I., C.A.S.I.) or by paper and pencil technique (P.A.P.I) and then stored in a database. In these situations, the amount of data to be coded plays a fundamental role in deciding which procedure can be adopted. In more details:

- for a large amount of data it is advisable to use AUC and subsequently CAC for the non coded cases;
- for a small amount of data and simple classification it is better to apply AUC;
- for a small amount of data, complex classification and high responses' variability it is more convenient to adopt CAC.

The following table summarises what stated before.

	Quantity of data/statements to be coded	
Classification structure	**Large number**	**Small number**
• **Simple**	AUC + CAC	AUC
• **Complex & high responses' variability**	AUC + CAC	CAC

Table 4: Coding activity after data collection

2.3. Quality measures for AUC and CAC procedures

The quality of the AUC procedure can be measured by two parameters: "*recall*" and "*precision*". The first one is the percentage of automatically coded verbal responses on the total ensemble of the examined ones and could also be indicated as the "coding rate". The grater is the computer capability in assigning codes directly, the larger is the cost reduction in using AUC, since the few descriptions left non-coded would require a small manual coding activity. The second parameter represents the percentage of automatically assigned correct codes (in comparison to those assigned by manual coding experts) on the total amount of coded descriptions. A good AUC procedure aims at maximising both parameters at the same time: this implies a continuous updating of the dictionary as well as the improving of the matching techniques. But the maximisation of the recall rate must take into account the precision rate and vice-versa:

- maximising the recall rate without considering the precision rate would mean to increase the number of assigned codes, worsening the overall quality;
- on the other hand, aiming at reaching the highest precision level would optimise the general quality but would drastically lower the number of successes (coded descriptions)

Therefore, to get a balanced performance of an automated coding procedure, the recall and the precision parameters must be always analysed and improved together.

Similar quality measurements could be used to evaluate the performance of the CAC procedure. The recall parameter has to be replaced by the "number of codes directly assigned by the coders", whereas the precision could be defined as before, that is percentage of correct assigned codes. The continuous update of the dictionary remains a basic condition for improving the overall quality, together with a periodic training for coders, especially for those working on complex classifications.

3. Coding systems selected by ISTAT

The needs of coding systems suitable for all ISTAT surveys induced to choose generalised coding procedures, that is, systems that were independent from the classification and the language used. This choice implied the construction of dictionaries for all the variables to be classified. The creation of the dictionary, that obviously precedes any other step of the coding process, is a very delicate phase since it has to make the official language close to the spoken one. This can be obtained by further elaborations of the dictionary as well as by its integration with empirical pre-coded responses:

- long and complex official descriptions containing more than one concepts are split in short and simple statements, expressing only one single concept and the same code is assigned to all of them;

- empirical responses provided by respondents during previous surveys or pilot studies are included in the dictionary in order to take into account the common way people use to express concepts (that, in general, is quite different from the official language).

Synthetically, for the construction of the dictionary the following elements need to be considered:

- the descriptions of the official classification associated with the relative codes;
- a large list of synonymous of the official descriptions in order to:
  - ⇨ make the dictionary close to the respondents' language as much as possible;
  - ⇨ restrict the coders' freedom in interpreting the verbal responses;
  - ⇨ make the coders' job easier: a big list of synonymous would probably foresee the major part of the possible verbal responses and this would definitely help manual operators in assigning codes even to very specific and peculiar responses.

ISTAT has chosen two systems for supporting the coding needs:

1. **ACTR v3** (Automated Coding by Text Recognition - Statistics Canada) for the implementation of the AUC procedure;
2. **BLAISE** (The Netherlands CBS) for the realisation of the CAC procedure.

3.1. ACTR v3 system

ISTAT has chosen ACTR basically because it is a generalised system, independent from the language and already used with success by other National Statistics Institutes. The basic logic of ACTR is inspired on the methods originally developed at US Census Bureau (Hellerman 1982) and uses matching algorithms studied by Statistics Canada researchers (Wenzowski 1988).

The coding activity is preceded by a quite sophisticated text standardisation phase, called "parsing", providing 14 different "parsing functions" (character mapping, deletion of trivial words, definition of synonymous, suffixes removal, etc…) able to remove grammatical or syntactical differences so that any two different descriptions, with the same semantic content, become identical. The parsed response to be coded is then compared to the parsed descriptions of the dictionary. If by this search, a perfect match is found, that is a "direct matching" is realised, then a unique code is assigned, otherwise the software runs an algorithm to look for the most suitable partial matches ("indirect matching"). In this latter case, the software takes out of the dictionary all the descriptions that have at least one parsed word in common with the input verbal phrase and assigns them a score, that is calculated as a function of the weight of each single word in common (the weight of each word is inversely correlated to its frequency of occurrence in the dictionary). The system orders the descriptions extracted from the dictionary by a descendent score rank and compares them with some user-defined threshold parameters. As a result the software returns:

- unique matches, when a unique code is assigned to a response phrase;
- multiple matches, when several possible codes are proposed;
- failed matches, when no matches are found.

While in the first case there is no need of human actions, the other two cases need to be evaluated by expert coders. This can be done by the "on-line" coding function: the operator enters the verbal description and the computer shows him the list of descriptions in the dictionary matching with it. Therefore ACTR provides the user either with a "batch processing" for coding and with an "on-line" function that represents a sort of "navigation" inside the dictionary. This "navigation" is quite simple in that the match of words in common between the input and the dictionary statements is realised by searching the dictionary in a flat way, from its very beginning to its very end. More articulated ways of "navigating" inside the dictionary are provided by BLAISE, as described in the following paragraph.

3.2. BLAISE system

The reason for choosing BLAISE can be found in its double functions:

- it is a system studied for the computer assisted data collection (C.A.T.I., C.A.P.I., C.A.S.I., C.A.D.I.);
- it also has a specific module for the assisted coding that realises a CAC procedure.

Therefore, BLAISE can realise the coding activity during the interview with all the resulting advantages (as previously described). Being a generalised coding system, it requires the construction of the dictionary for each classification; as soon as a dictionary is loaded, BLAISE system gives it a tree structure on the basis of the classification codes. This structure takes into consideration the hierarchy among codes, assigning tree-branches to the highest code levels (generalised descriptions) and tree-leaves to the most inferior code levels (detailed descriptions). An example of this structure is the classification of the "*Time use activities*":

0	**Personal cares**	**(branch)**
	01 To sleep	***(sub-branch)***
	011 To be sleepy	*(leaf)*
	012 To lay on bed for sickness	*(leaf)*
	02 Food and drinks	***(sub-branch)***
	021 Meals	*(leaf)*
1	**Working activity**	**(branch)**

After this first step, BLAISE searches for codes, offering three different modes of "navigation" inside the dictionary:

1. a stepwise coding or a tree coding;
2. a dictionary coding;
3. a mixture of stepwise and dictionary coding.

1) Stepwise coding or tree coding. At first, the operator opens a window showing only the highest levels of the classification (0=Personal Cares, 1=Working activity, etc.); he chooses the corresponding branch and then a more detailed menu appears showing the sub-branches of the selected category. The process continues, showing menus with more and more specific descriptions, until the final code is found.

2) Dictionary coding. The operator enters the input description and the computer searches for it inside the dictionary: if the description is found, than the computer extracts it and assigns it the corresponding code automatically. If the description is not found, than the machine performs matches even for parts of words, that is, for two or three consecutive letters – the so called "*digrams*" and "*trigrams*" - at the beginning or the end of the word. This means that the computer is able to perform a wider search, thus offering the operator a greater possibility of assigning a correct code. The extracted descriptions (each representing a dictionary's record) are listed in a descending order according to the number of matches realised for each word in a record.

3) Mixture of stepwise and dictionary coding. The operator starts with a stepwise coding until he is able to select the branch. Then he switches to the dictionary coding and the computer shows him a list of descriptions where he makes a textual search.

4. ISTAT experiences

For the major part of ISTAT surveys, the coding activity was realised at the end of data collection and was decentralised and carried out by local municipal employees. Since this kind of organisation was very time-consuming and did not guarantee a high quality level of the results (De Angelis R., Macchia S. and Mazza L. 2000), it was decided to test the automated coding procedure. Moreover, the approaching of the Census Surveys (Population Census and Industry Census) amplified the need of an automated system, making the manual coding quite unsuitable, because of the huge amount of data to be coded and the non-availability of a large number of coders for the assisted coding activity.

4.1. The experience with ACTR

The initial experience of automated coding was realised using ACTR. The first step of the process was the construction of the dictionary for each variable involved in the coding procedure. The variables entering the coding activity, the amount of descriptions of the corresponding official classifications and the total number of records of each coding dictionary are summarised in the following table.

	Official Classification items	**Dictionary Descriptions**
Variables		
Industry	1.208	23.239
Occupation	7.072	14.510
Educational level	133	3.202
Country (nationality)	194	1.931
Municipality	8.100	59.876
Enterprises' legal status	28	105
Pathology	12.000	187.000

Table 5: Coding environments built in Istat for AUC

As this table shows, the number of dictionary's descriptions is always much greater than the number of the classification's items. This is due to the operations made on the dictionary descriptions, such as further elaborations of the official texts, and to the inclusion of empirical responses (as described above) necessary to improve the recall rate of the automated coding system. The automated coding of the above variables was carried out several times as they were observed in different surveys, as indicated in the following table.

Variables	**Surveys**
Industry	Population Census (PC) 1991 – Quality Survey
	Labour Force Pilot Survey 1999
	Population Census (PC): 1° pilot survey 1998, 2° pilot survey 2000
	Intermediate Industry Census 2000 (Long & Short Form)
Occupation	Population Census (PC) 1991 – Quality Survey
	Health Survey (data collection 1994)
	Labour Force Quarterly Survey - 1998
	Labour Force Pilot Survey - 1999
	Population Census (PC): 1° pilot survey 1998, 2° pilot survey 2000
Educational level	Population Census (PC) 1991 – Quality Survey
	Population Census (PC): 1° pilot survey 1998, 2° pilot survey 2000
Country (nationality)	Population Census (PC) 2° pilot survey 2000
Municipality	Population Census (PC) 2° pilot survey 2000
Enterprises legal status	Intermediate Industry Census 2000 (Long Form)
Pathology	Causes of death survey - 1999 (test)

Table 6: Surveys in which AUC was adopted

The quality of ACTR system was measured by the *recall* (R) and *precision* (P) rates. The results are summarised in the following prospect.

% Values	VARIABLES													
	Occupation		Industry		Educational level		Country		Municipality		Enterprises legal status		Pathology	
SURVEYS	R	P	R	P	R	P	R	P	R	P	R	P	R	P
PC Quality survey	72,5	90,0	54,5	-	86,6	99,7								
Health survey '94	72,3	97,0												
Labour Force '98	72,0	97,3												
IIC (Short Form)			47,0	-										
IIC (Long Form)			58,8[1]	-							94,0	100		
Labour Force pilot '99	66,7	99,9	43,5	84,8										
PC 1° pilot survey '98	65,5	98,1	51,2[2]	93,7	75,7	99,7								
PC 2° pilot survey '00	68,8	96,8	51,9	90,0	87,0	99,0	83,2	100	94,5	100				
Causes of death survey													82,0	100

Table 7: AUC results in terms of Recall and Precision rates

[1] Average value obtained during survey activities. The recall rate grew up to 70% after the training phase of ACTR coding environment

[2] The recall rate for "Industry" variable shows a higher level for the Industry Census (Long Form survey) than for the Population Census: this is due to the difficulty the households faced in answering the question not quite easy to understand

The above data show that while the precision rate is high (generally higher than 90%) for all variables and surveys, the recall rate varies a lot.

- High precision rate indicates a satisfactory validity of the automated coding system (when a code is assigned, then it is assigned correctly), that is, in the case of ACTR, a well-structured dictionary and a good working parsing phase.

- High recall rate, on the other hand, could not be uniquely defined because its level strictly depends on the complexity of the classification and on the responses' variability. The above table shows that, in general, the recall rate is lower if the classification is complex and a high level of variability in the responses' wording is verified. As a matter of fact:

    - recall rate reaches the highest point - approximately 94% - for the "*Municipality*" and *"Enterprises Legal Status",* variables that have a simple classification structure (a tree-structure with only main branches) and a limited wording variability;

    - it then slightly goes down to 87%-82% for *"Pathology"*, *"Country"* and *"Educational level"* variables, because their simple classification structure (a bit less simple for the Educational level) is combined with a quite high responses' variability;

    - it finally reaches the lowest levels – 70%-50% on average - for the *"Occupation"* and *"Industry"* variables, where the very complex classification is joined with a high variability in the responses.

Obviously, recall rate can always be improved: this would imply a continuous training for the automated procedures, that is the updating of the dictionaries as well as the refining of the matching techniques.

4.2. The experience with BLAISE

The experience with the CAC procedure is not so developed yet, since all C.A.T.I. and C.A.P.I. surveys have been carried out by external companies (C.A.S.I surveys are still at their beginning). Besides, the need of a coding system based on a different approach from that used by ACTR came out quite recently in conjunction with the *"Time use"* survey that aims at studying how time is spent. This survey is based on a sample of 25.000 Italian households that have to fill in a diary, for a certain period of time, describing the activities they perform during the day, specifying them in detail every quarter of hour (Camporese R., Ranaldi R., 2001). C.A.D.I. technique (*Computer Assisted Data Input*) will be used, therefore the coding activity will took place at the end of data collection. The implementation of the coding process for this survey will require many efforts since the classification of the time use activities is very complex and respondents' language presents a high variety of descriptions to express the same action. For example, the same activity can be expressed using different verb tenses: "*I prepare breakfast*", "*I prepared breakfast*" or "*I'm preparing breakfast*". Moreover, this richness of the language is combined with a high degree of tolerance for ambiguities in that vague expressions, as well as sentences with several meanings, are always used and understood in the spoken language (Camporese R., Ranaldi R., 2001). These elements generate many difficulties for the coding process because the "language of codes" presents an opposite structure: it is rigid, does not accept ambiguous concepts and requires univocal interpretation for definitions.

To face these difficulties, it has been decided to use BLAISE, whose characteristics make it suitable for the coding of time use activities.
As a matter of fact:

- The possibility offered by BLAISE of matching parts of words (*trigrams* and *digrams*) would overcome the problem of different verb tenses: the match will be realised using the root of the verb, thus avoiding the influence of different tenses in code assignation. This implies that there is no need of having a standardisation procedure to eliminate the verbs endings (as realised by ACTR *parsing phase*).

- Moreover, common misspellings and word abbreviations have less impact on matching. In fact, while for ACTR a correct word and a misspelled one are considered as completed different, the BLAISE matching procedures, through *trigrams* and *digrams,* makes it possible to avoid this problem.

- It is possible to insert in the dictionary comments or notes, to be used by coders as a "help on line" to solve ambiguities.

- Finally, as already said, BLAISE supports three different ways of "navigating" the dictionary which eases and speeds up the coders' job.

The quality level of BLAISE performance will be evaluated as soon as data from "Time use" survey will be available.

5. Conclusions

Automated coding seems to be essential by now, for both cost and time reduction and to obtain standardised results with a high quality level. How deciding which is the most suitable automated system to adopt depends on the users' specific needs, although some general factors, like the survey technique, the amount of data, the interview length, the structure of the classification and the variability of the verbal responses, could be used as a guide in this choice. A combination of specific needs and general factors must always be made in order to help users in taking the right decision. The systems chosen by ISTAT have given good results even if other studies will be performed to improve them, to set standards on the methodologies for building coding dictionaries and even to test alternative solutions.

References

Appel M. and Hellerman E. (1983). Census Bureau experience with Automated Industry and Occupation Coding. In American Statistical Association, *Proceedings of Section on Survey Research Methods*, pages 32-40.

BLAISE for Windows 4.1 Developer's Guide – 1999.

Camporese R. and Ranaldi R. (2001). Time Use Activities: Translation from Sentences to Codes. In CLADAG2001, Meeting on the Classification and Data Analysis Group of the Italian Statistical Society, Palermo, Italy.

Chen B., Creecy R. and Appel M. (1993). Error control of automated industry and occupation coding. *Journal of Official Statistics*, vol. 9: 729-745.

Cochran W. G. (1977). *Sampling Techniques, 3rd ed..* Wiley, New York.

De Angelis R., Macchia S. and Mazza L. (2000), Applicazioni sperimentali della codifica automatica: analisi di qualità e confronto con la codifica manuale, Istat Quaderni di ricerca – Rivista di statistica Ufficiale, 1, 29-54

Dumicic S. and Dumicic K. (1994). Optical reading and automatic coding in the Census '91 in Croatia. In *Conference of European Statisticians, Work Session on Statistical Data Editing*, Cork, Ireland 17-20 October, Working Paper n. 2.

Everitt B. S. (1977). *The Analysis of Contingency Tables.* Chapman and Hall, London.

Hellermann E. (1982). Overview of the Hellerman I&O Coding System. Internal document, US Bureau of the Census, Washington.

Lyberg L. and Dean P. (1992). Automated Coding of Survey Responses: an international review. In *Conference of European Statisticians, Work session on Statistical Data Editing*, Washington DC.

Kalpic D. (1994). Automated coding of census data. *Journal of Official Statistics*, vol. 10: 449-463.

Knaus R. (1987). Methods and problems in coding natural language survey data. *Journal of Official Statistics*, vol. 1: 45-67.

Massingham R. (1997). Data capture and Coding for the 2001 Great Britain Census. In *XIV Annual International Symposium on Methodology Issues*, 5-7 November, Hull, Canada.

Tourigny J.Y. and Moloney J. (1995). The 1991 Canadian Census of Population experience with automated coding. In United Nations Statistical Commission, Statistical Data Editing, 2.

Wenzowski M.J. (1988). ACTR – A Generalised Automated Coding System. Survey Methodology, vol. 14: 299-308.