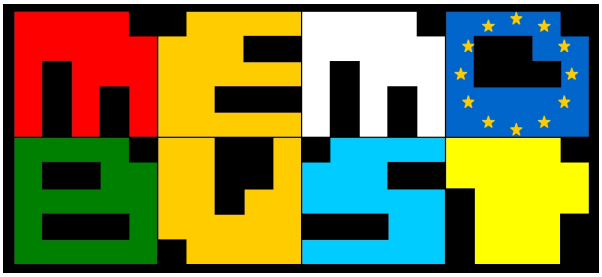

This module is part of the

Memobust Handbook

on Methodology of Modern Business Statistics

26 March 2014

Theme: How to Build the Informative Base

Contents

General section

1. Summary

Coding of verbal responses of a statistical survey could be defined as assigning numeric codes to statements according to a manual of official classification. Performing this activity manually is costly, time consuming and error prone, so the computer support for this process is increasing. The informative base to be used for this purpose is the fundamental part of any computerised approach, because it must fulfil at least two requirements: make the computer rely on knowledge similar to that inside the human mind and, starting from the content of the official classification manual, process and enrich it with selected descriptions and/or synonyms derived from empirical responses given in previous surveys, so as to make the language closer to the spoken one. Logical steps to be carried out to build informative bases are described, as well as particular aspects to be taken into consideration either when coding is done in a completely automated way or with human support.

2. General description

Coding of verbal responses of a statistical survey could be defined as assigning numeric codes to statements according to a manual of official classification. The knowledge concerning these official classifications is usually contained in the classification manuals which describe, using appropriate words, the meaning of each concept and its code, providing definitions, specific details and exceptions.

When the coding process is done manually, the coder must be trained on using the classification and on how to find the information in the manual to assign the correct code corresponding to the textual response. But, as confirmed by the experience of a lot of NSIs, manual coding is time consuming, costly and error prone, so computer support for this activity is desired. It is also evident that when the computer is used, it must be given additional information from human experts lacking from the classification manual. As a matter of fact there are at least two aspects a human mind can consider while reading that a computer cannot, if not trained:

* the grammar and syntax rules (singular/plural, masculine/feminine, verbs declinations, …);
* semantic knowledge (the computer does not know the real meaning of words, it does not know, for instance, that an *orange* is a *citrus fruit*).

For these reasons a computer tool, in order to be used to code text responses, should:

* have an informative base supplying the computer knowledge similar to that from the human classification experts;
* have a search engine able to perform a text standardisation so as to identify a word independently from all the variable parts of it.

Both these aspects are even more important in two situations:

* when the coding process is made automatically, with a batch procedure and without any human intervention;
* when the coding process is made with computer support and directly by the respondent (in self-administered interviews).

As a matter of fact, in the first situation the computer must reason as a human mind does, because it has no other input at its disposal apart from its informative base and the responses to be coded. In the second situation it must be considered that the respondent, differently from the interviewer, is not an expert of the classification and might not know technical words used in manuals.

Regarding the informative base, this is the fundamental part of any computerised approach for coding. It is mainly constituted of a *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 the coding procedure to 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.

As far as the search engines for text processing are concerned, they can be more or less sophisticated, but it must be considered that text responses of statistical surveys to be coded according to official classifications are generally not too long and usually do not consist of very complex syntax constructions.

Several studies have been made at NSIs to identify or develop suitable tools to process texts in order to perform coding (Lyberg and Dean, 1992): 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 algorithms for automated coding software build 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. Other dictionary algorithms include in the dictionary a description belonging to the expert-coded file 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.

Other coding systems use a 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*"[1]. 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 - perform 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 sophisticated instruments use the so-called "*artificial intelligence*". One of these is the "Connection Machine" – Thinking Machine Corp. – that is a computer working with thousands of

[1] The system mainly used in Istat in several surveys, ACTR -Automatic Coding by Text Recognition, produced by Statistics Canada (Wenzowski, 1988), is based on a weighting algorithm. The new release of ACTR is called GCode.

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 (Appel and Hellerman, 1983).

Whichever coding system is adopted, the problems faced when building these dictionaries are the same: first of all the official classification manual must be transformed so as to be 'processable' by computerised systems and then lots of sources must be integrated in order to make it closer to spoken language used by respondents. As a matter of fact, textual descriptions of classifications are designed for manual coders, who assign codes making deductions and referring to their specific knowledge of the matter or to their personal cultural background (Knaus, 1987). A 'processable' dictionary, on the contrary, should include only **synthetic**, **analytical** and **unambiguous** descriptions (while two or more different descriptions can be associated to the same code, the same description must never be associated to different codes).

In general, the following logical steps can be defined in the activity of construction of dictionaries (D'Orazio and Macchia, 2002):

- **Simplifying descriptions** → often a description which summarises more than one concept is associated to a single code, while the typical respondent is used to refer to a single concept (for example: the Istat classification on Occupation assigns a single code to '*mathematicians and statisticians*', while the respondent will presumably answer only '*mathematician*' or '*statistician*', according to his specialisation). In these cases, it is necessary to split the phrase in two or more descriptions and to associate each of them to the same code.

- **Defining synonyms** → classifications contain generic words relating to categories, while people answer using specific words (for example, the Economic Activity classification considers the '*production of cereals*', while the respondent might answer only '*production of wheat*' or '*production of corn*'). Here it is necessary to list all the specific synonymous words to which the generic word refers to.

- **Eliminating exception clauses** → automated coding software do not usually reason in terms of exclusion, so they cannot understand the meaning of 'apart from…' and of similar clauses used to exclude certain categories from the class. In this case, it is necessary to take the 'apart from…' away and to verify that the 'excluded' concepts are included in other classes.

- **Treating open classes** → classifications usually include open descriptions, that is 'Other …' which means 'other than the concepts already specified' (for example: '*Other specialised clerks*', where different kinds of specialised clerks have already been listed in the preceding classes). Also in this case it is necessary to list all the explicit descriptions to which the open description refers. To make the list as complete as possible, it is advisable to use the responses given in previous surveys which have been coded as 'Other…' by expert coders.

- **Integrating with reference material** → the 'processable' dictionary can be usefully widened with descriptions coming from other related classifications. For example, each time the classification of Economic Activity has an element regarding the production of a certain '*category of products*' (summarising an implicit list) the specific classification of products can be used to enumerate the explicit list of products.

- **Integrating with empirical responses** → official classifications texts are often not very similar to the way people speak and their updating is slower than real world changes. Thus it is advisable to include in the dictionary selected descriptions derived from empirical responses, given in previous surveys, that had already been coded by classification experts.

As the dictionary is extended according to these criteria, its performance will increase, especially in the case of automated coding.

As far as Istat's experience is concerned, for instance, an application to code with the Economic Activity classification has been set up since 1998 and was updated following the new classification releases and used for several surveys. The informative base was built starting from the classification manual and enriched through the analysis of results of its use in each survey. As a matter of fact, after using this application to code the textual responses of a survey, non-coded responses were examined to find the cause, e.g., an ambiguity in the original text or the response contained some synonyms not present in the informative base. In the latter case, the missing synonym was added to the informative base.

In order to give an idea about the impact of the size of the informative base on the results of automatic coding, the dictionary grew from 27,306 descriptions to 34,180 and the average percentage of coded texts (on the total number of texts to be coded) from 50% to 71% in the last surveys (Macchia, Murgia, and Vicari, 2010).

Finally, it must be mentioned that, when the coding process is made interactively with the computer support (during the interview, either by the respondent of by the interviewer, or after the interview by coders), computer tools often provide functions to navigate inside the informative base.

When the classifications have a hierarchic structure it is advisable that the software tools allow to navigate inside the dictionary according to the classification tree.

Blaise for instance, developed by Statistics Netherlands, manages navigation in its coding database according to three different methods:

- through the textual matching → the respondent/coder enters the text and Blaise extracts from the database the texts which have one or more trigrams in common with the keyed text;
- according to the classification tree → the respondent/coder selects the classification branch of the highest level and then goes deeper in the sub-branches towards the lower levels;
- with a mixed method → the respondent/coder selects the classification branch of the highest level and then enters the text to perform textual matching among descriptions belonging to the selected branch.

3. Design issues

4. Available software tools

5. Decision tree of methods

6. Glossary

For definitions of terms used in this module, please refer to the separate "Glossary" provided as part of the handbook.

7. References

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

BLAISE for Windows 4.5 Developer's Guide (2002).

D'Orazio, M. and Macchia, S. (2002), A system to monitor the quality of automated coding of textual answers to open questions. *RESEARCH IN OFFICIAL STATISTICS (ROS)*, N.2.

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

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

Macchia, S., Murgia, M., and Vicari, P. (2010), Integration between automatic coding and statistical analysis of textual data systems. Journée d'Analyse des Données Textuelles JADT, Rome.

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

Interconnections with other modules

8. Related themes described in other modules

1. Coding – Main Module

9. Methods explicitly referred to in this module

1.

10. Mathematical techniques explicitly referred to in this module

1.

11. GSBPM phases explicitly referred to in this module

1. Sub-process 5.2 Classify and code

12. Tools explicitly referred to in this module

1. BLAISE for Windows
2. GCode, new release of ACTR -Automatic Coding by Text Recognition

13. Process steps explicitly referred to in this module

1.

Administrative section

14. Module code

Coding-T-Informative Base

15. Version history

Version	Date	Description of changes	Author	Institute
0.1	20-07-2012	first version	Stefania Macchia	Istat (Italy)
0.2	21-11-2012	second version (following first revision)	Stefania Macchia	Istat (Italy)
0.3	17-01-2014	third version (following EB review 08-01-2014)	Stefania Macchia	Istat (Italy)
0.3.1	21-01-2014	preliminary release		
1.0	26-03-2014	final version within the Memobust project		

16. Template version and print date

Template version used	1.0 p 4 d.d. 22-11-2012
Print date	21-3-2014 18:05