Computer Assisted Coding by Interviewers[1]

Wim Hacking, John Michiels & Saskia Janssen-Jansen (Statistics Netherlands)

1. Introduction

Coding activities play an important role in a statistical production process. These activities are usually associated with the assignment of responses to predefined codes in a classification, so that these responses become available for further data-editing operations. Coding can be done at various stages of statistical production: during data collection by respondents or interviewers or during the data-editing process by coding experts and/or automated systems. In most cases coding is a rather trivial exercise: when questions are clear and unambiguous and the number of answering categories is limited an answering category can easily be provided by respondent or interviewer. The situation becomes more difficult when the number of answering categories is large or several open text answers are required. Examples of these more difficult coding activities are the assignment of a respondent's educational background or occupation to a corresponding classification. In these examples usually more than one question is involved in gathering the information required for assigning valid codes and part of these questions are of the open text type.

In previous years the approach adopted at the *division of social statistics* for coding (open text) responses from a number of related questions was the following. First, the answers to these questions are collected using CAPI or CATI modes of data collection. In the case of long answers they are interpreted and modified by the interviewer to obtain a concise response. At the statistical office the collected and sometimes edited information is then fed to a 'batch' process for automated coding. The records that are not coded are classified interactively by coding experts in a second pass. Coding economic activities at the *division of business statistics* was partly done manually at Statistics Netherlands and partly externally, at the chambers of commerce.

In this paper we would like to describe alternative coding techniques that are more successful than the traditional approach. The technique chosen depends on a number of factors (and possibly more than given here):

- The desired level of detail for statistical output (publications, international obligations) and for intermediate processes (e.g. weighting).
- The desired quality of coding (error rates).
- The available budget and expertise for development, implementation, operation, and maintenance.

The first point is relevant because for obtaining low detail classifications in most cases less elaborate methods can be developed that produce good results. The desired quality puts similar restraints on the choice of coding method. As higher demands on detail and accuracy require more complex coding techniques, it will in general also mean higher development costs.

In this paper we will discuss an alternative coding strategy that has been developed and selected for implementation at Statistics Netherlands: Computer Assisted

[1] The views expressed here are those of the authors and do not necessarily reflect the position of Statistics Netherlands

Coding by Interviewers (CACI). The new strategy is more cost effective than traditional approaches with comparable level of detail and reliability. The available expertise for implementing this approach can usually be found at a statistical office; if not, the required effort to obtain this knowledge is certainly less than for some more advanced methods in machine learning or for expert systems. In the next section we will describe the methods that have been applied for semi-automatic coding. In section 3 we will discuss the interactive coding approach as applied in the interview process at Statistics Netherlands, based on the methods from section 2. Finally, section 4 shows results using these new techniques. Here the percentage of coded records and coding reliability for each technique will be considered, along with results on interview lengths for CACI (based on field work done by the division of social statistics). In addition, some early results for the coding of economic activity at the division of business statistics are given.

2. Coding Techniques Used

We will start by giving an overview of the coding techniques used. The merits and drawbacks of each technique will also be discussed in brief. Basically, three coding techniques have been chosen depending on the (coded) material at hand: in 2.1 we discuss the situation that (many) previously coded records are available in electronic form. In 2.2 we start with a registry containing an extensive description for each code. In 2.3 no electronic material is available and a search file is constructed specific for the coding process. Depending on what initial material is at hand methods have selected for semi-automatic coding.

2.1. Using previously coded material

A number of techniques for automatically classifying text strings has been developed and implemented by the Institute for Knowledge and Agent Technology (IKAT) at Maastricht University: Nearest Neighbours (NN), Term Frequency Inverse Document Frequency (TFIDF), and Naïve Bayesian (NB) (Smirnov 2003, Kaptein 2005). All of these techniques involve learning algorithms that need a training data set with combinations of text strings and corresponding codes for 'optimization' or 'training'. The nature of this optimization depends on the technique being considered. Training a learning algorithm produces a text-classifier with an approximate mapping of text strings to codes. The text-classifier assigns a weight or probability to each code in the classification and the code with the largest weight or probability is then selected.

The problem is to determine the reliability of this type of classification. There are ways to do this using another kind of learning-machine called meta-classifiers. These classifiers use 'meta-information' from text-classifiers: in case the code assigned by the text-classifier equals the true code in the training data set the combination of text string and code vector is labelled as 'good', otherwise it is labelled as 'bad'. The meta-classifier uses this kind of information in order to optimize a set of rules for correctly assessing text-classifier results: with these rules the meta-classifier decides whether the codes assigned by the text classifier for new text strings are to be considered as 'good' or as 'bad'. There are different ways in which these rules can be constructed and hence there are different meta-classifiers. Examples of meta-classifiers are described in the thesis of Kaptein 2005 and the main results are mentioned in section 4.2.

Instead of coding the material afterwards based on classifier techniques we can also apply them during the interview in the field. We will an example to clarify the approach chosen. Suppose that we are interested in coding the open text 'carpenter' (as a possible answer to the question: what is your current occupation?). This answer has been given in previous surveys many times before and at the statistical office it has been assigned a number of different occupation codes (using additional information on job activities). It is now possible to calculate conditional probabilities $P(Code_i|\text{ 'carpenter'})$ representing the frequency with which each code has been assigned to 'carpenter'. Therefore to each word in the open text answer a vector of conditional probabilities can be determined. In case the open text contains more than one word the vectors are added to produce a combined vector of weights (not probabilities). For example, in the open text answer 'carpenter at shipyard' there are probability vectors for 'carpenter', 'at', and 'shipyard'. These vectors are added to produce a vector of weights:

Carpenter: $P(code_1|\text{ 'carpenter'})=0,60$; $P(code_2|\text{ 'carpenter'})=0,20$; .....
At: $P(code_{14}|\text{ 'at'})=0,02$; $P(code_2|\text{ 'at'})=0,01$; $P(code_{11}|\text{ 'at'})=0,01$; .....
Shipyard: $P(code_2|\text{ 'shipyard'})=0,50$; $P(code_4|\text{ 'shipyard'})=0,35$; ..... +

Carpenter + at + Shipyard : $Weight(code_2)=0,71$; $Weight(code_1)=0,60$; .....

Or in a formula:

$$Weight(Code_i)=\sum_j P(Code_i\,|\,Word_j)$$

The probability vectors for the individual words are stored in index files containing the conditional probabilities $P(Code_i|Word_j)$. After calculating the weights for each code based on the search string, the code descriptions of codes with the largest weights are then presented to the respondent if the combined weight of the first (say) 6 codes is above a certain threshold. The respondent then selects a particular code description and a code is uniquely established. In case the combined weight is too small there are too many codes possible for the given answer and these can not all be presented; in that case it is better to ask additional information in order to limit the number of possible code descriptions.

The main advantage of using text- and meta-classifiers is that they represent a cost-effective technique as far as operational and maintenance costs are concerned. Another advantage is that the use of these classifiers does not require a deep understanding of the classification problem at hand. Accurate training data are not necessary although "noisier" data degrades the classifier performance. However, there are some drawbacks: the cost of developing text- and meta-classifiers can be high. And often the expertise needed for development and implementation is not available at a statistical office. Moreover, a lot of data are needed to train text-classifiers and the technique does not increase our knowledge about the coding process itself.

2.2. Using registries

The most cost effective approach to coding is probably by using results that have already been collected elsewhere. For example, if one is interested in coding the economic activity of one-man businesses in the Netherlands one can link the complete files of the population and business administrations using the social security number as the connecting key. Although in principle registries could work well for coding purposes one usually faces the following problems: the registry does not exist (yet), the registry does exist but is not available to the statistical office, the quality of the linking field(s) or the target variables is insufficient, or the information is outdated. In the latter case there is still need of a questionnaire and the respondent should be asked whether the (outdated) information is still valid. Formulating appropriate questions is not always easy. Also one has to link sampling frame and registry in order to identify sampling elements for which the information in the registry is applicable.

On can also use the registry as a search file which contains information about the item to be coded. For example, in the case of businesses, we use a search file that consists of records containing information about business names (legal and trade names) and the corresponding codes for economic activity. The search strategy is then to compare open text answers with text strings in the search file and select records that ‘match’. Records in the search file ‘match’ with open text answers if the words in the record description are very similar or equivalent to the words in the open text answer. For example, a respondent may have indicated that he or she is an employee at ‘Philips lighting’. This answer is very similar to the record description ‘Philips lighting bv’ in the search file for business names. One would also like to consider other records with similar descriptions like ‘Philips lights nv’ or ‘Philips displays’. The similarity between descriptions can be measured using a kind of metric (i.e. a kind of distance formula). We applied the ‘Levenstein metric’ where the difference between descriptions is measured by counting the minimum number of character inserts+deletes+changes that are required to go from one string to the other. Records with descriptions that are similar to the open text answer are presented to the respondent for further selection.

This kind of approach is only applicable if we are looking for terms that the respondent knows relatively accurately (in this case business name) and the existence of a registry that maps those terms to the desired code (SBI in this case). In addition, the quality of the codes assigned depends on the quality of the codes present in the registry.

2.3. Using handmade search files

If no material is available to that links open text to the codes, a search file may be constructed. For each code a definition must be made and stored in conjunction with the code. Using this naïve approach will result in very little hits, since the descriptions given by the respondents may differ greatly. In order to alleviate this problem a synonym file was added that translates “field terms” (as used by the respondents) to “standard terms” (as defined in the search file).

Coding education at the division of Social and Spatial Statistics

This approach was chosen for the coding of education, where a search file is used containing a list of the most frequent regular educations along with their defining characteristics; a tiny extract is shown below to illustrate the concept:

Table 1: For each code a number of characteristics have been determined and these are stored in separate column in a search file. Each code corresponds with one record. With each column a specific question is associated.

Level	***Field***	***Teacher***	***University Degree***	***Teacher Type***
Medium level	English	No		
Higher level	English	No		
University	English literature	No	Master	
Higher level	Interpreter English	No		
Higher level	Translator English	No		
University	English	Yes	Master	First degree
Higher level	English	Yes		Second degree

To code education, the first question asked concerns the name of the education. When the respondent’s answer has been entered (e.g. “higher level English”) the search engine tries to locate matching records. This results in a preliminary selection of educations. This selection is usually too large to present to the respondent and therefore further questions about the education’s defining characteristics are asked in order to limit their number. The possible answers for subsequent questions consist of the distinct items from column X corresponding with the question. If the set of selected records has finally become sufficiently small (less than 7 educations) then either one of two situations can occur: 1) only one education is left; this education is presented to the respondent for confirmation, or 2) 2-6 educations are left; the interviewer presents these options to the respondent for final selection. Occasionally, more than one record left, but all questions have been asked. In this case, the code must be derived manually at Statistics Netherlands based on the answers and selections form the respondent.

Coding economic activity at the division of Business Statistics

However, for the coding of economic activity at the division of business statistics this approach was not sufficient. Apart from a much broader range of field terms, which can be solved by extending the synonym file, another problem needed to be addressed. In order to explain this problem we must first describe the search file in more detail:

Table 2: an excerpt of the search used to code economic activity.

(SBI) Code	***Description of the code***	***Process***	***Good***
X	Cultivation of agricultural product	Cultivation	agricultural product
…	…	Cultivation	…
X+N	Cultivation of flowers	Cultivation	flowers
Y	Processing of potatoes	Processing	potatoes
Z	seedpotatoa wholesale	wholesale	Seedpotatoa
Z+1	Potatoes for consumption wholesale	Wholesale	Potatoes for consumption
…	…	Wholesale	…
Z+M	Wholesale of computers	Wholesale	Computers

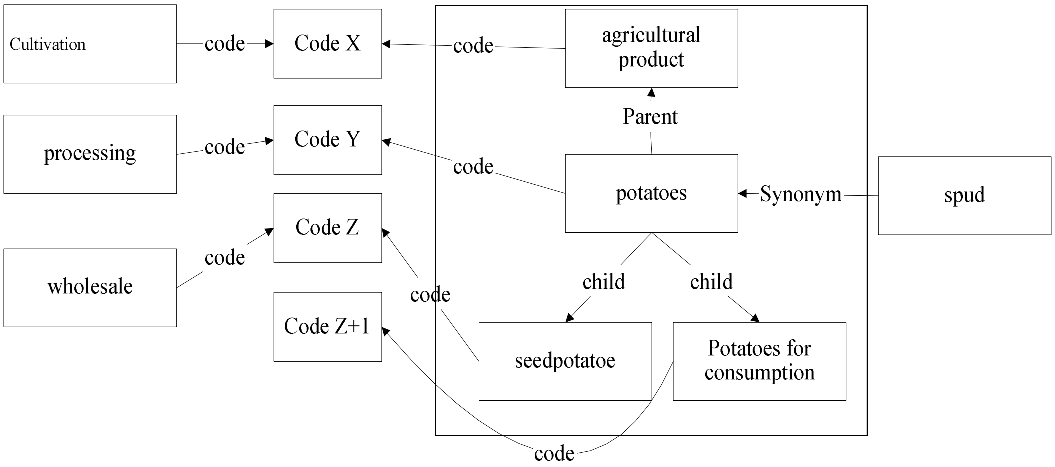
Based on the excerpt of the search file above we will describe the search result for 3 search strings as shown in table 3 in the first column. The second search works fine. The first search string specifies the good too narrow; as a result one only gets hits for the word *cultivation* (i.e. codes X….X+N). The same holds for search string 3.

Table 3: Three prototypical search strings to illustrate the necessity to introduce classifications in the search method.

No	***Search term***	***Comment***	***Hits without classification***	***Hits with classification***
1	Cultivation of potatoes	Just hits through *Cultivation*	N	1
2	Processing of potatoes	Both words match	1	1
3	Potatoes wholesale	Just hits through *wholesale*	M	2

In order to know the level of detail when specifying the good, one needs to know the code, but this is not known in advance, by definition. To solve this problem a classification has been added for goods and code descriptions. A very small extract of the classification relevant for the current problem is shown below inside the box as part of the semantic network.

Figure 1: This network is based on the three file types (search file, synonym file and classification file) which have been merged into one. The specific roles of the files have been translated into edge-types (code, parent, child and synomym).



Search in this network proceeds as follows:

1. For each word of the search string assign a score 1 to the corresponding node
2. Pass the activity on to the next nodes (following the arcs in the direction) with one restriction in the classification part: once a child edge has been passed, parent edges can no longer be traversed and vice versa.
3. Perform step 2 until "code nodes" are encountered.

This kind of search is called spreading activation (see Crestani 1997, Berger 2004). The spreading of activation in our implementation is governed by the formula:

$A_x = \sum A_i * EdgeStrength_{x,i}$

A_x : the activity of node ***x***

A_i : the activity of the nodes ***i*** *that have node* ***x*** *as input*

$EdgeStrength_{x,i}$: the "strength" of the edge from node ***x*** *to node* ***i***

In the current implementation $EdgeStrength_{x,i}$ = 1.

All three search strings in table 3 contain "potatoes". Due to the classification this word is expanded to more general terms ("agricultural product") and to more narrow terms ("seedpatatoa" and "Potatoes for consumption'). As a result we already have an activity 1 for the codes X, Y Z and Z+1. If we search with the other term from the search string the relevant code(s) will receive an activity or score 2 and the others will remain at a score 1.

This approach has been used successfully for the initial search after an open text has been given by the respondent. If too many records remain for direct selection an approach similar to that for coding education is used, i.e. additional questions are asked to reduce the number of records.

The disadvantage of this method, compared with the previous two ones, is the considerable amount of work needed to construct the search file, the synonym file and the classification files. However, it does explicit our knowledge and understanding of the classification itself.

3. Computer assisted Coding by Interviewers (CACI)

3.1. Concept of computer assisted coding

At Statistics Netherlands we developed and implemented an interactive technique for the classification of economic activities, occupations, and educations, where the coding activities are mainly performed by interviewers during CAPI-interviews. In this process the interviewers collect open text answers from respondents and enter these text strings on their laptop computers. The laptop software then evaluates the entered text strings and generates a small list of appropriate codes. Next, these codes are presented to the respondent in the form of a short list of code descriptions from which the respondent can select a unique code description. Having done this a code has been established. The role of interviewer and laptop software is to guide the respondent to a correct answer. Open text answers for which no code can be found are referred to coding experts at the statistical office.
The CACI methods described in section 2 have been implemented as laptop software with search engines and search files. The implementation of these methods in questionnaires will be discussed in the next paragraph, where we will consider the mechanism of the search strategies in detail.

3.2. Coding economic activity for the labour survey

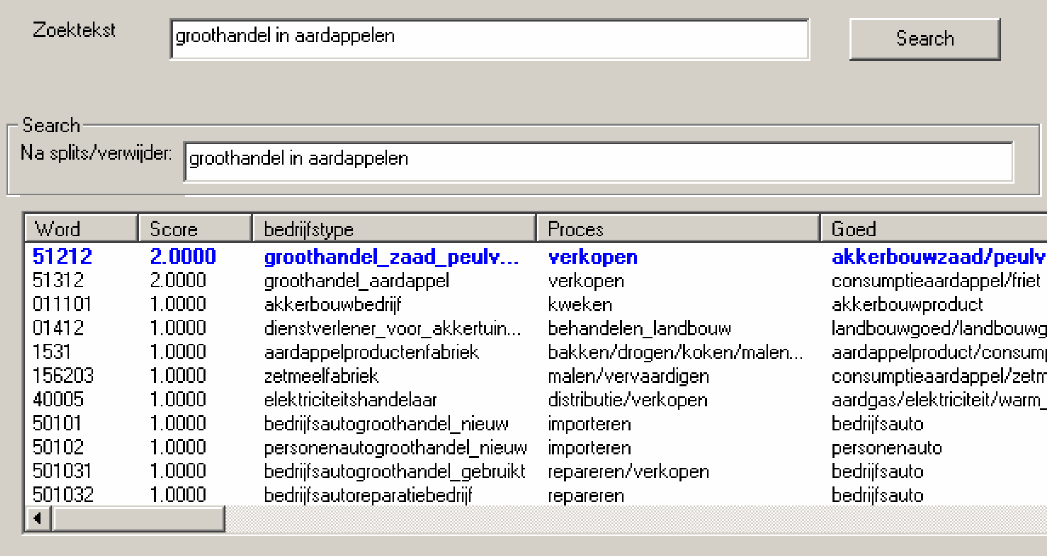
For the coding of economic activities of businesses two search files are used, based on methods from sections 2.1 and 2.2. One of these files is a registry: the General Registry of Businesses. It contains statistical information for most active businesses located in the Netherlands including their legal names, trade names and a code number representing the main economic activity. Because of the size of this registry only a small number of businesses are selected in the search file. The selection is based on business size (businesses with 100 or more employees are selected). The reason for this selection is that a large part of the employees work in a relatively small number of large businesses; businesses with 100 or more employees represent approximately only 1% of the population, however approximately 57% of all jobs are found within these businesses (Statistisch Jaarboek 2004, Statistics Netherlands). The second search file is based on results from previous surveys. It contains coded descriptions of economic activities as has been discussed in 2.1.

The first question in the Blaise questionnaire (concerning business variables) is about business size. The answer to this question determines which search file is going to be used: large businesses can probably easily be located in the registry data. Smaller businesses are not present in this file and in this case the search engine should use the search file with historical data. If a business has 100 or more employees the question ‘What is the name of the entire business?’ is presented to the respondent. The answer is entered and the search engine then tries to locate appropriate business names in the search file, using the Levenstein metric for calculating the ‘closeness’ of words. If more than one hit and less than 7 hits are found the selection is presented to the respondent in order to identify a single business name and establish a unique code. In case the search engine cannot find similar business names or if more than 7 business names are found the respondent is asked to describe the business activities. This is also the first question for businesses having fewer than 100 employees. When an open text answer has been entered the search engine tries to locate records with similar descriptions of business activities in the search file containing historical data (based on the method from 2.1). If there are fewer than 7 hits or if there are 6 hits which have sufficient weight the selection is presented to the respondent in order to select an appropriate economic activity. In case no hits or too many hits are found (or if no option is applicable) the search engine interface is switched off and the routing of the questionnaire returns to the Blaise-environment. Here a hierarchical type of question is asked to determine economic activity with less detail (1 or 2 digits instead of 3 digits).

3.3. Coding economic activity at the division of economical statistics

The prototype for coding economic activity at the division of business statistics is currently implemented as a standalone desktop application. Once started, the application asks the user to describe the main economic activities of their business. If there is a single code that has the largest score, the description is considered to be coded.

Figure 2: Initial screen of the prototype to code economic activity. The search string is "groothandel in aardappelen" (= "Potatoes wholesale"). As described in section 2.3 this leads to 2 best candidates.



Zoektekst: groothandel in aardappelen

Search

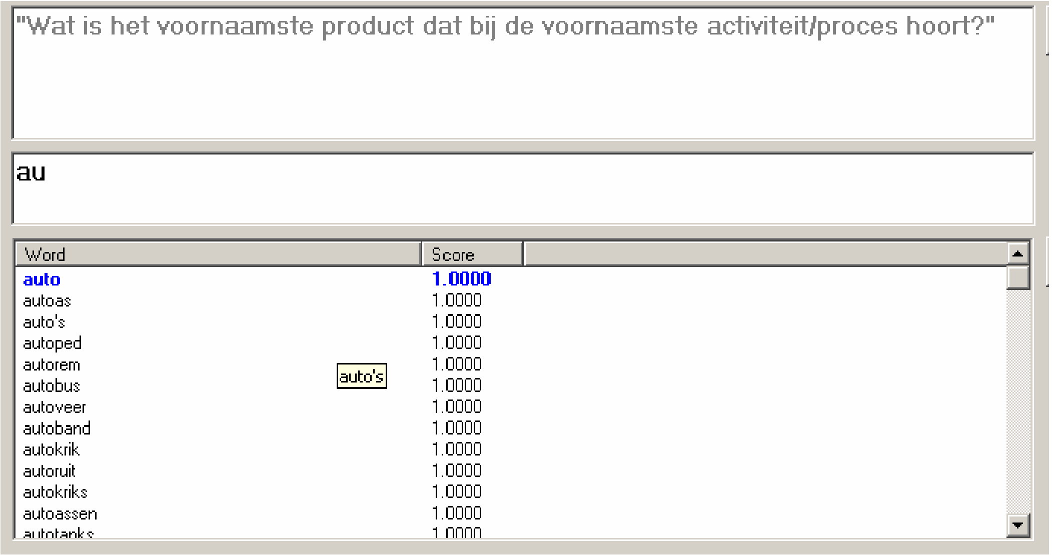
Na splits/verwijder: groothandel in aardappelen

Word	Score	bedrijfstype	Proces	Goed
51212	**2.0000**	**groothandel_zaad_peulv...**	**verkopen**	**akkerbouwzaad/peulv**
51312	2.0000	groothandel_aardappel	verkopen	consumptieaardappel/friet
011101	1.0000	akkerbouwbedrijf	kweken	akkerbouwproduct
01412	1.0000	dienstverlener_voor_akkertuin...	behandelen_landbouw	landbouwgoed/landbouwg
1531	1.0000	aardappelproductenfabriek	bakken/drogen/koken/malen...	aardappelproduct/consum
156203	1.0000	zetmeelfabriek	malen/vervaardigen	consumptieaardappel/zetm
40005	1.0000	elektriciteitshandelaar	distributie/verkopen	aardgas/elektriciteit/warm_
50101	1.0000	bedrijfsautogroothandel_nieuw	importeren	bedrijfsauto
50102	1.0000	personenautogroothandel_nieuw	importeren	personenauto
501031	1.0000	bedrijfsautogroothandel_gebruikt	repareren/verkopen	bedrijfsauto
501032	1.0000	bedrijfsautoreparatiebedrijf	repareren	bedrijfsauto

If not, more questions will be asked by the application to try and reduce the set of possible codes as shown in the screenshot in figure 3.

If, after asking some questions, the set of possible codes is less than 7, the respondents asked to choose from a fixed list of descriptions.

Figure 3: Screenshot from the prototype. At the top the question ("what is the main good used for the main activity ?"). Below that the answer can be typed; while entering letters, the most fitting distinct items from the column goed (=good) will be displayed below.



"Wat is het voornaamste product dat bij de voornaamste activiteit/proces hoort?"

au

Word	Score
auto	**1.0000**
autoas	1.0000
auto's	1.0000
autoped	1.0000
autorem	1.0000
autobus	1.0000
autoveer	1.0000
autoband	1.0000
autokrik	1.0000
autoruit	1.0000
autokriks	1.0000
autoassen	1.0000
autotanks	1.0000

3.4. Technical implementation

The coding system as implemented at the Division of Social and Spatial Statistics consists of the following components:

- The main search routines have been implemented in C++ because of speed

- Some additional logic and the graphical user-interface have been implemented in VB

- Finally, the above two components are embedded in a Blaise questionnaire, so that interviewers can start the search engine at specific questions using a key stroke.

For the Division of Business Statistics a prototype has been implemented based on the same C++ code in combination with some new VB code to render a desktop application.

4. Results

An important part of the development of automated coding and the implementation of computer assisted coding is the evaluation of their performance. There are two issues to address here. First, there is the question of effectiveness and reliability, i.e. is the percentage of cases that are coded sufficiently high and is the reliability and level of detail according to specifications? The second issue is about the requirements for the new technique to work well in a production environment. This covers matters like: can interviewers handle the new type of questionnaire, what additional training is required, what other software applications are needed for maintenance and monitoring, etc. In this paper we will only address the first issue, although the second one is very important once one decides to move beyond the development phase.

Definitions

Evaluation of the performance of CACI and text-classifiers will be based on a number of parameters: coverage and error rate. The definition of these parameters for CACI is as follows:

Coverage = [# cases that are coded / total # cases] x 100%.

Error rate = [# cases that are not coded correctly / # coded cases] x 100%.

where a 'case' is one occupation, one business, or one education for which a code is to be established (# means: 'number of').

The qualification of a case as coded correctly or not in CACI is made by comparing results with the codes from coding experts. A code is classified as correct if the CACI code is equivalent to the result established by the coding expert or if this latter code was present among the selection made by the search engine. In other cases the CACI code is classified as false. For text-classifiers the error rate is calculated differently. It is determined from the accuracy of the meta-classifier:

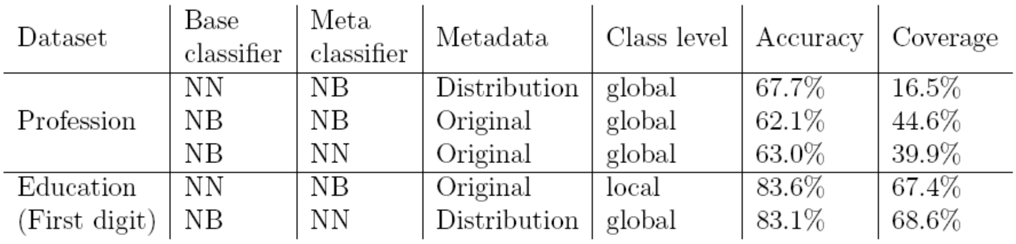
Accuracy = [# cases correctly labelled 'good' / total # cases labelled 'good'] x 100%

The error rate of the meta-classifier is then simply: 100% - *Accuracy*.

Performance of text-classifiers

The meta-classifier approach from section 2.1 was applied two datasets from Statistics Netherlands, occupation (6 digits) and the level of education (1 digit). The training and test data were obtained from several Labour Force Survey files, containing 182038 cases. Optimal settings for different parameters have been determined, e.g. text representation and classification algorithms. The best results for the two datasets are given in table 1.

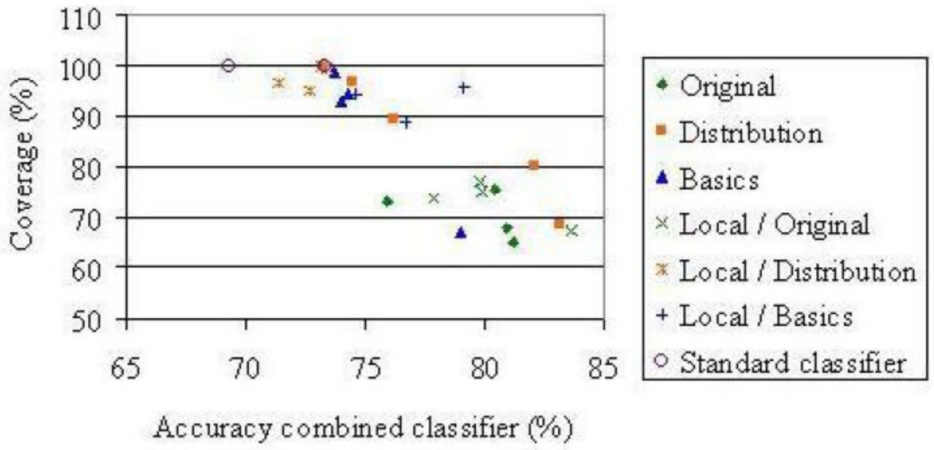
Table 4: Accuracy and Coverage of the meta-classifier approach when coding occupation (SBC codes) or education level. (*NB: Naïve Bayesian; NN: Nearest neighbours)*



Dataset	Base classifier	Meta classifier	Metadata	Class level	Accuracy	Coverage
Profession	NN	NB	Distribution	global	67.7%	16.5%
	NB	NB	Original	global	62.1%	44.6%
	NB	NN	Original	global	63.0%	39.9%
Education (First digit)	NN	NB	Original	local	83.6%	67.4%
	NB	NN	Distribution	global	83.1%	68.6%

As apparent from figure 4 a choice must be made between coverage and accuracy as they are inversely related:

Figure 4: Accuracy and Coverage of the meta-classifier approach when coding education level. The descriptions of the methods mentioned can be found in Kaptein 2005.



One the conclusions from her thesis confirmed the starting point of the work described in this paper, namely that the largest problem encountered with automatic coding is the (lack of) quality of the input data. The input contained much "noise", e.g. misspellings and descriptions that were either too vague or ambiguous.

4.3. Performance of CACI

In this paragraph and in paragraph 4.4 we present results on coverage and error rates for the coding techniques discussed previously in this paper. The performance of the CACI coding technique has been measured in two phases. The first phase consisted of a field experiment (in September 2003) with approximately 700 respondents using a Labour Force Survey questionnaire with CACI modules for coding economic activity, occupations and education, the second phase included an analysis of the production version of the Labour Force Survey in January 2004 with the same CACI-modules as in the first phase test. In this case a regular survey sample was used with 11607 respondents. The first phase was intended to establish preliminary results on coverage and reliability of the CACI coding technique before release of the production version in 2004. In the second phase the coverage was measured anew in regular production. Also, an additional check of coding reliability has been realized for a limited number of respondents. At the time of writing this article results on error rate were available for the coding of economic activity (based on a sub sample of 500 respondents). In table 5 an overview of these results is given.

Table 5: results from the coding in the field for 2 periods.Code$_{field}$ coded correctly if same as code Code$_{manual}$ from coding experts, or Code$_{manual}$ was present among the selection made by the search engine, but not selected by the respondent.

Attribute	First phase: September 2003		Second phase: January 2004	
		%		%
Number of respondents	699	100	11501	100
respondents age: 14 years or older	564	81	9128	79
respondents with a job	364	52	5726	50
Number of businesses to be coded	383	100	5978	100
coded with search engine	315	82	4653	78
coded with hierarchical question	49	13	947	16
not coded	19	5	378	6
Number of occupations to be coded	364	100	5726	100
coded with search engine	289	79	4299	75
not coded	75	21	1427	25
Number of (current) educations to be coded	107	100	1650	100
coded with search engine	88	82	1235	75
not coded	19	18	415	25
Number of completed educations to be coded	1199	100	19127	100
coded with search engine	989	83	15569	81
not coded	210	17	3558	19
Error rate economic activity of businesses		<12		7
Error rate occupations		<10		-
Error rate educations		<13		-

4.4. Performance of the semantic network

Although not used in the field yet, there are some early results for the coding of economic activity at the division of business statistics (based on section 2.3):

Number of codes found:

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	0	26	1,9	1,9	1,9
	1	1174	85,1	85,1	87,0
	>1	180	13,0	13,0	100,0
	Total	1380	100,0	100,0	

The results can be divided into 3 categories:

0 hits: the word combination in the search string is not found in any of the descriptions of economic activity in the search file.
1 hit: either exactly one hit of there is just one hit with the highest score
>1 hits: there is more than one record with the highest score

The results for the complete coding process (including the additional questions) are not available yet at the time of writing.

4.5. Length of interview

In addition to coverage and accuracy, interview lengths have also been measured during the pilot at the division of social and spatial statistics. Interview lengths are an important factor to consider when one is interested in making a more cost-effective coding process using computer assisted coding in the field. If interview lengths increase then more interviewers are needed to carry out the same number of interviews. The gain in cost-efficiency due to a reduced number of manual coding activities may then be partly or completely cancelled due to increases in interview lengths. Therefore we also present results on interview lengths for the Labour Force survey in 2003 (with the old questionnaire) and in 2004 (with the new CACI questionnaire). When compared to the old situation the interview lengths of CACI questionnaires on economic activity of businesses, occupations and education change considerably, but not all in the same direction. The average interview length for businesses decreases whereas the average interview lengths for occupations and educations increase. This is illustrated in table 6 below.

Table 6: since the primary drive of the project was a reduction of fte's, too large an increase of interview time is unwanted. The average time per person for each of the 3 subjects education, occupation and business. After an initial increase, the total time for the 3 subjects returns the value for the old situation.

Questionnaire	**Interview length (old)** Labour force survey 2003	**Interview length CACI** Labour force survey January-March 2004		
		January	February	March
Business	72 sec	53 sec	48 sec	49 sec
Occupation	36 sec	49 sec	46 sec	47 sec
Education	177 sec	209 sec	196 sec	190 sec
Total	285 sec	311 sec	290 sec	286 sec

From the table it becomes clear that the interview lengths of the new CACI questionnaire decrease over time. In January 2004 the interview length for businesses, occupations, and educations combined was on average +26 seconds larger per interview when compared to the situation in 2003. In February and March of 2004 the combined interview length was only slightly larger than in 2003: on average respectively +5 seconds and +1 second per interview. This may indicate that interviewers get accustomed to the new questionnaire.

The change in interview lengths has a plausible explanation. In the old questionnaire information on business names, location and type of business activities were collected in order to establish the code for economic activity. A large part of this information, however, was redundant. In the CACI questionnaire this redundancy is removed and often a single question (concerning name or type of business) is sufficient for a successful classification. The situation is different in the case of occupations and educations: here the number of questions required to establish a code is not reduced and interview lengths actually increase. The reason for this to occur is that each question in the new questionnaire demands extra tasks to be performed. Apart from posing a question and entering an open text answer a code has to be selected from a listing produced by the search engine. For this selection process to work the interviewer must present these codes to the respondent and the respondent has to identify an appropriate option. The number of questions that include this search-and-select strategy is small in the case of occupations (approximately 1-2) and large for educations (on average >3 as all educations of the respondent are recorded). Therefore the increase in interview length is largest with educations.

5. Conclusions

Interactive coding during interviews has an important advantage over traditional approaches. If the information contained in the open text answers is not sufficient to select a unique code the interviewer can ask the respondent to limit the selection. In case the information is sufficient no extra questions need to be asked. This possibility of getting extra feedback from respondents where required is an improvement on techniques that do not use an interactive scheme.

Although a large part of the coding effort is transferred to the field interviewers, measurement of the timing of the field interviews shows that, after an initial rise in interview time, there appears to be no extra time required; in addition, interviewers appreciate the feedback given by the search engine.

The development of CACI requires programming expertise and knowledge of statistical processes. This experience is usually available at a statistical office. Development and implementation include the construction of a questionnaire, software tools, test procedures, and the writing of documentation. Also maintenance of software and search files for production is important. Examples are the updating of search files due to changes in classifications, the correction of errors in search engines, and transmitting new versions of search engines and search files to interviewers.

Based on our results we conclude that CACI is a promising technology for the classification of survey data. Most of the occupations, educations, and economic activities can be coded successfully giving results with sufficient reliability. In addition, it is cost efficient in operation and search files can be constructed and updated automatically using new data from coding experts. This makes it also a technology with low maintenance costs.

6. References

Berger H., Dittenbach M., Merkl D., *An Adaptive Information Retrieval System based on Associative Networks*, APCCM04

Conrad, F. (1997), “Using Expert Systems to Model and Improve Survey Classification Processes”, *Survey Measurement and Process Quality*, John Wiley & sons, inc.

Crestani, F. (1997), *Application of spreading activation techniques in information retrieval,* Artificial Intelligence Review, 11(6):453–582

Kaptein R. (2005), *Meta-Classifier Approaches to Reliable Text Classification* , master’s thesis, IKAT, Maastricht University.

Mitchell, T. (1997), *Machine Learning*, McGraw-Hill.

Smirnov, E.N., Sprinkhuizen-Kuyper, I.G., Wiesman, F.J., Donkers, J., Postma, E.O., and van den Herik, H.J. (2003), *Typing professions adequately*, IKAT, Maastricht University.

Statistics Netherlands (2004), *Statistisch Jaarboek 2004*, published by Statistics Netherlands.