Interactive coding of economic activity using trigram search in BLAISE III

Omar S. Hardarson, Statistics Iceland

Abstract

This paper presents the results of a study at Statistics Iceland on the feasibility of using trigram-coding to code economic activity. The results show that the total success rate, i.e. the rate of producing acceptable codes, of trigram-coding and residual manual coding can be expected to range between 89.1% and 97.7%, depending on the classification level and type of population. The total time of such coding is less than the total time of exclusively manual coding. The paper also shows that while there was no difference in the coding success between experienced and inexperienced interviewers, the efficiency of trigram-coding is mainly determined by the quality of the index of descriptions.

1. Introduction

Any survey of the population can be seen as a systematic codification of verbal data from respondents. Answers are sometimes coded as ayes or nays but more often than not the response categories are more varied than that. In computer assisted interviewing, however, it is not advisable to exceed more than one screenful of alternatives for any one question. The step towards a more complex structure of alternatives is nevertheless only a question of degree, albeit highly dependent on the available software algorithms as well as coding dictionaries.

Versions 2.0 to 2.4 of the BLAISE CAI system provided facilities for complex classifications with hierarchical coding or alphabetical search in lists. Version 2.5, and later BLAISE III 1.1, offer trigram-coding in addition to the former two.

Roessingh and Bethlehem (1993) compare the three types of coding facilities provided by BLAISE when coding articles in the Dutch Expenditure Survey. They conclude that at least for this subject matter the trigram-coding is better than the alphabetical, but hierarchical coding is by far the best. They point out, however, that their coders were very experienced, knowing many of the codes by heart, as well as the fact that hierarchical coding is largely dependent on the availability of a “workable classification of items “ (p. 132). They also find that trigram-coding can be successful even if the coders have no experience with it.

With regard to trigram-coding Roessingh and Bethlehem (1993) point out that its success is determined by the quality of the dictionary or list of descriptions at hand. The coding file for trigram-coding, however, does not have to be as long as the list for alphabetical coding where one needs to have multiple entries of the same items in a different word order. The importance of the coding file is emphasised by others, e.g., Lyberg and Dean (1992), Bushnell (1993) and Martin (1993).

Lyberg and Dean (1992) offer the dictum that "[a]utomated coding and residual manual coding should be less expensive than the same operation being 100% manually coded" (p. 23). This may not hold true for interactive coding. Bushnell (1993) lists 13 characteristics for a computer assisted coding system. An "ideal" system should include as many of these as possible. Cost effectiveness is only one of these criteria.

In theory computer assisted coding should be better than manual coding of data generated by the same interviews. This is because the interviewers can solve any ambiguities by asking the respondent. The manual coder is on the other hand restricted by the information already recorded. This advantage of interactive coding is, however, offset by other factors. The interviewers may not be sufficiently trained, they have less resources than the manual coder and the time to code may be unacceptably long (see Martin 1993). The decision to use interactive coding has thus to take all these factors into account.

From 1992 Statistics Iceland has conducted its Labour Force Survey as well as other surveys with the help of the BLAISE CAI system. Versions 2.3x, 2.4 and 2.5 were used until early spring 1996, when the agency introduced BLAISE III 1.1 in its survey work. Until 1996 the search and look-up facilities in the BLAISE system were used very sparingly and almost exclusively for the coding of labour unions in the Labour Force Survey. This was primarily due to the difficulty of the BLAISE 2.x system in handling non-ASCII characters in strings. This drawback was corrected in the BLAISE III version. In April 1996 we introduced trigram-coding for coding the municipality of the local unit and the ISCED level of completed further education in the LFS. Both of these proved to greatly facilitate the interview process. We were therefore keenly interested in knowing how the trigram-coding module would perform when coding complex concepts like economic activity of local units.

The study of the efficiency of trigram-coding was carried out in two separate steps. The first step was part of a Pilot Survey for the Small Business Survey in the summer of 1996. The second step consisted of an experiment within the November 1996 Labour Force Survey. The first two parts of the study were aimed primarily at estimating the reliability of the trigram-coding vis-à-vis manual coding, whereas the third part focused on examining the relative cost of this type of coding. The Small Business Survey as well as the Labour Force Survey could also be used to assess the results of the Pilot Survey.

2. Design of the study

The Small Business Pilot Survey (SBPS). The first part of the study was based on a pilot survey for the Small Business Survey in July 1996. The

purpose of the pilot survey was partly to examine the feasibility of coding economic activity of the local unit during the interview by using the trigram-coding facility in BLAISE III 1.1.

A random sample of 400 businesses was drawn from a part of the Small Business Survey population. Of these seven were no longer part of the frame, and 26 others were dropped from the sample for other reasons. There were thus 367 individuals in the final pilot survey sample.

The Pilot Survey was conducted on 11, 12 and 14 July between 17.00 to 22.00. Seven interviewers were hired to conduct the survey, four on the first two days and then three in addition on 14 July. The list of descriptions was taken directly from the index of the ISAT-95 manual, with 3,831 entries.

Contact was made with 243 individuals, of which twelve did not have any business activity in the reference period (the year 1995), leaving 231 cases to be coded.

The Small Business Survey (SBS) was primarily aimed at assigning the new ISAT-95 codes[9] to the activities of unincorporated businesses. The total population of unincorporated businesses exceeded 28 thousand. The majority of these had ISIC[10] codes assigned, which could be directly matched with ISAT-95 codes. A total of 8,059 businesses were coded in the survey. The file of coding descriptions from the Pilot Survey was used for trigram-coding with the addition of professions which could be matched with an economic activity code, along with some other minor changes. The total number of entries was 4,222 coding descriptions.

The assigned codes from either the manual or the interactive coding were checked by a separate coder. If this coder found any ambiguity he would look closer at the data and make necessary changes to the codes. In some instances the code was left unchanged. In this study the results of the review process are used to determine the acceptability of the initial activity codes.

The Labour Force Survey (LFS). The second part of the survey consisted of an experiment within the November Labour Force Survey 1996. The Icelandic LFS is carried out twice a year, in April and November. It is a panel survey with some 4,400 participants with a net response rate of 90%. Approximately three quarters of the respondents are carried to any subsequent survey. The panel design enables the technique of dependent interviews in which the answers from a previous survey are confirmed or changed in the interview. In the survey a total

---

[9] The ISAT-95 classification is the Icelandic equivalent of the NACE (rev. 1) classification of industry, where the first 4 digits are more or less compatible with international standards while the fifth digit is exclusively Icelandic.

[10] A three digit code based on the international standard, ISIC, rev. 1, which had been in use at Statistics Iceland since the late sixties.

1,774 cases were previously recorded and confirmed, leaving 1,227 cases to be coded afresh.

The experiment was designed so that half of the respondents answering the question on economic activity of the local unit in their first job were randomly assigned to two groups:

The responses of the first group were recorded by the interviewers, and then the resulting text was used for the initial search in the coding dictionary using the trigram-coding utility in BLAISE III 1.1. If the interviewers did not succeed in finding a relevant code they were asked to record a further description of the business activity in a separate field of the type OPEN.

The second group, or the control group, was only given the question on the main functions of the local unit, which was then coded by an experienced coder, using an MS Access form to review all relevant information provided by the respondent (such as age, gender, education, name of the firm, place of activity, occupation and employment status) as well as access to the Icelandic Business Register and the ISAT-95 dictionary of descriptions. The expert coder also coded the cases where the interviewers failed to assign a code.

In the LFS economic activity is coded to the level of the fourth digit, even if the required precision level need only correspond to the first two digits. Nevertheless the same list of descriptions and five digit codes were used in the LFS as in the SBS. Out of 23 interviewers 10 had participated in the SBS. These were designated as experienced interviewers, even if many of the remaining interviewers had previous experience with the LFS. The total number of cases which were coded either manually or interactively was 1,227.

The inexperienced interviewers were given the same type of training in coding economic activity as the experienced interviewers. The training consisted of a short lecture on the concepts and structure of the ISAT-95 classification frame and hands-on practice, totalling four hours prior to the survey.

The procedure of reviewing and editing the activity codes in the SBS as described above applied also to the LFS.

3. Accuracy of trigram-coding vs manual coding

The assignment of ISAT-95 codes in the SBPS was made with the trigram-coding facility of BLAISE III 1.1. On completion of the coding the respondents were requested to confirm whether they had been allocated the correct code. In addition, the interviewers recorded the relevant information on the activity in a field of the type OPEN. An experienced coder was subsequently given the task of coding the activities afresh without having access to the results of the trigram-coding or earlier ISIC codes.

The two resulting codes were then compared. If the codes were identical, we assumed that they were acceptable. If the codes differed, the author

then reviewed the codes and coded them as acceptable or unacceptable. If the recorded data was too scant in order to code with precision to the fifth digit the former ISIC code, if available, was used to assess the acceptability of the result.

Table 1 shows the results of the review. Using the data recorded by the interviewers, the expert coder could assign 222 codes out of 231 or 96.1% at the five digit level, while the interviewers assigned 213 codes or 92.2%. All of the cases where the expert coder failed to assign a code were also the cases where the interviewers did not record sufficient information to assess the code without resorting to the existing ISIC codes. The acceptance rate of the interactive coding and manual coding at the five digit level is 95.1% and 97.7%, respectively.

Table 1. Results of the interactive and manual coding. SBPS 1996.

	Interactive coding			Manual coding		
	5 digits	4 digits	2 digits	5 digits	4 digits	2 digits
Acceptable	196	197	205	217	220	227
Unacceptable	10	9	5	5	5	2
Insufficient information	7	7	3	0	0	0
Not coded	18	18	18	9	6	2
Total number of cases	231	231	231	231	231	231
Hit rate[1]	92.2	92.2	92.2	96.1	97.4	99.1
Acceptance rate[2]	95.1	95.6	97.6	97.7	97.8	99.1
Success rate[3]	87.7	88.2	90.0	93.9	95.2	98.3

[1] The number of assigned codes over the total number of cases
[2] The number of accepted codes over the number of assigned codes
[3] Hit rate times acceptance rate

In an actual interview environment the cases that cannot be coded interactively will be referred to manual coding. The above results indicate that trigram-coding and residual manual coding produce a total success rate of approximately 95.1% at the five digit level, 95.6% at the four digit level and 97.7% at the two digit level,.

Table 2 shows how the 1996 Small Business Survey performed. In 91% of the cases the interviewers succeeded in finding an ISAT-95 code, of which 96.4% where considered acceptable after checking for errors. The acceptance rate was somewhat higher if the requirement was only to code to the two digit level, or 97.9%. These results are in line with the results of the pilot survey.

Table 2. Hit rate and acceptance rate using trigram search while interviewing. SBS 1996.

	ISAT-95 5 digits	ISAT-95 4 digits	ISAT-95 2 digits
Hit rate	91	91	91
Acceptance rate	96.4	96.6	97.9
Success rate	87.7	87.9	89.1
Number of cases	8,059	8,059	8,059

In the November 1996 Labour Force Survey the hit rate was similar to the hit rate in the 1996 SBS (Table 3). The acceptance rate was, however, considerably lower, or 88.7% at the four digit level and 92.7% at the two digit level. These result indicate that the total success rate for interactive and residual manual coding amounts to 93.1% at the two digit ISAT-95 code, as compared to 97.7% in the SBPS. At the four digit level the total

success rate in the LFS can be estimated as 89.1%, as compared with 95.6% in the SBPS.

The difference between the experienced and inexperienced interviewers was not statistically significant, in fact the success rate was almost equal (Table 3). The results seem, however, to indicate that the inexperienced interviewers were more inclined to resort to guessing if they could not find an appropriate code immediately.

Table 3. Hit rate and acceptance rate using trigram search while interviewing. LFS November 1996.

	ISAT-95 4 digits	ISAT-95 2 digits		
	Total	Total	Inexperienced interviewers	Experienced interviewers
Hit rate	92.5	92.5	93.8	90.5
Acceptance rate	88.7	92.7	91.6	94.3
Success rate	82.0	85.7	85.9	85.3
Number of cases	610	610	357	253

The fact that there was no difference in the success rates between experienced interviewers and inexperienced, indicates strongly that one has to look elsewhere for the determinants of the success rate. Table 4 shows that the interactive coding performed very differently depending on the section of economic activity of the edited ISAT-95 code. In both the LFS and SBS these variations were statistically significant.

Table 4. Success rate of trigram-coding by economic activity, 4-digit level. SBS 1996, LFS November 1996.

Percentages	SBS 1996	LFS Nov. 1996
Agriculture	95.3	94.4
Fishing	96.3	100.0
Manufacturing	76.8	76.8
Electricity and water supply	–	…
Construction	89.9	92.5
Wholesale and retail trade, repairs	85.5	75.3
Hotels, restaurants	88.2	92.9
Transport, storage and communication	95.1	77.8
Financial intermediation	90.5	84.6
Real estate, renting and business activities	85.8	73.9
Public administration	…	71.4
Education	61.9	86.2
Health and social work	92.9	89.6
Other services	88.8	73.6
Total	87.9	82.0
Number of cases	8,050	610
χ^2	324.10	28.61
df	12	13
p	.000	.007

4. The cost of trigram-coding vs manual coding

The length of the interview process was measured in the November LFS by using the inherent time measurement in the CATI environment, whereas the length of the manual coding and post-editing process was measured simply by noting the time of opening and closing the relevant forms in MS Access. This proved to be a crude measurement. In both cases a better procedure would have been to measure the time it took to process individual cases.

An experienced coder carried out the manual coding, while the review and editing were handled by an inexperienced coder who only had knowledge of the basic concepts and structure of the ISAT-95 classification but no prior practical experience. This coder was given insufficient instructions so he did not always close the forms while taking lunch hours or working on other things. For the purposes of this paper the idle time that could be verified was eliminated.

Table 5 shows the main results of the length of interview measurements. The control group (manual coding) has a shorter average length of interview than the interactive group. This difference is statistically significant at the .085 level. When the interactive interviews are broken down by the success of the trigram-coding, the "no-hits" stand out as considerably prolonging the interview. The difference between the "no-hit" average length of interview and other "interactive" interviews is significant at the .005 level. Even though the success rate is independent of the experience of the interviewers, Table 5 shows clearly that the inexperienced interviewers were far more likely than the experienced ones to spend time in a futile search in the trigram-coding dictionary.

Table 5. Average interview time in seconds by method and success of coding economic activity. LFS November 1996.

Averages		Length of interview in seconds		
	N of Cases	Total	Inexperienced interviewers	Experienced interviewers
Manual coding	617	443	485	375
Interactive coding	610	469	513	407
Not coded	46	601	777	439
Not acceptable[1]	64	474	501	417
Acceptable[1]	500	456	495	402
Total	1,227	456	498.6	392

[1] 4-digit level

We would, of course, expect that the average length of interviews would increase when adding an interview item to the questionnaire. On average the number of interview items, i.e., the number of recorded questions, trigram-codes and remarks, was 1.2 items more for the interactive group than the control group. As shown in Table 6 the effects of interactive coding disappear when controlling for the number of interview items.

Table 6. ANOVA Interview length by Manual/Interactive coding with Interview items

		Unique Method				
		Sum of Squares	df	Mean Square	F	Sig.
Covariates	Interview Items	11.314.590	1	11.314.590	193,7	0,00
Main Effects	Manual/Interactive coding	43.271	1	43.271	0,7	0,39
Model		11.515.136	2	11.515.136	98,6	0,00
Residual		71.505.736	1224	71.505.736		
Total		83.020.872	1226	83.020.872		

Table 7 shows the results of the measurements of time consumed while coding manually and while reviewing and editing both the interactive and manual codings. In addition, the estimated time of coding in BLAISE is taken into account, using the difference in average interview time between interactive coding and manual coding in Table 5.

Table 7. Time used in manual and interactive coding of economic activity. LFS November 1996.

4-digit level	Manual Coding		Interactive Coding		Total
	N of Cases	Hours	N of Cases	Hours	hours
Coding	663	6.13	610	4.33[1]	10.46
Checking	663	2.60	565	2.22	4.82
Editing[2]	34	1.36	73	2.92	4.28
Total		10.09		9.47	19.56

[1] Estimated time.

[2] In a number of cases the editing resulted in unchanged codes.

Table 8 shows the estimated time of interactive and residual coding for all 1,227 cases, using the hit rate estimates in Table 3 and the time measurements in Table 7. The table also shows the estimated time if the cases were exclusively manually coded. The table shows that one can expect that the time of actual coding will decrease by 21% while the time of editing will increase by the 39%. The total coding time using interactive coding as a primary method of coding is 8% less than if manual coding were used exclusively.

Table 8. Estimated time of coding economic activity using trigram and residual manual coding.

4-digit level	Manual Coding		Interactive Coding		Total	Excl. manual
	N of Cases	Hours	N of Cases	Hours	hours	hours
Coding	92	0.85	1135	8.06	8.91	11.34
Checking	92	0.36	1135	4.46	4.82	4.82
Editing	5	0.19	83	3.31	3.50	2.52
Total		1.40		15.83	17.23	18.68

Source: Tables 3 and 7

We represent in this paper the cost estimates only in terms of time consumed. Analysis of the pay structure at Statistics Iceland revealed that the average cost per hour of interviewer work is approximately the same as the average cost per hour of a professional with a university degree. This is because the interviewers worked predominantly night and weekend hours, and the fact that university education in Iceland is relatively poorly remunerated (Institute of Economic Studies 1997).

5. Discussion

In this study we have established that using interactive or trigram-coding as a primary means of assigning economic activity codes would produce an total success rate ranging from 89.1 to 97.7%, depending on the classification level and type of survey. This is in line with other studies on the accuracy of this type of coding (ILO 1990). In the post-editing process the total success rate can be increased substantially.

The results of the SBPS indicated that the interviewers were less successful than the manual coder in assigning ISAT-95 codes to the economic activity. The failures of the manual coder to assign a code was primarily due to imperfect data recorded by the interviewers.

In the SBS the trigram-coding performed as expected from the results of the pilot study. This was not the case for the LFS, where the success rate of the interactive coding was by far lower than expected. There were at least four significant differences between the two surveys to account for this.

First, in the SBS the respondents were asked a specific confirmation question, which was not the case in the LFS.

Second, in the SBS the time factor was of no importance, as each interview only covered a handful of questions. Many of the interviewers preferred to have the ISAT-95 manual available in order to be able to code the economic activities with greater accuracy.

Third, the respondents from the two surveys did not come from the same population. The SBS population is by far better informed of the functions of their businesses than the LFS population, which for the most part consists of employees with often only a sketchy knowledge of the functions of the firms with which they are employed.

Fourth, the two populations differed in the fact that the SBS population was more homogeneous than the LFS population. The economic activities of the small business owners tended to concentrate in certain sectors of the economy, and they were more likely than the firms in the LFS to cover only one possible ISAT-95 code.

The interesting result of this study is that there is no difference between experienced and inexperienced interviewers with regard to the success rate of trigram-coding. There was a slightly greater tendency for the inexperienced interviewers to have a higher hit rate but this was offset by more unacceptable codes, indicating that they may resort more to guessing when they do not find an appropriate code. These differences are, however, not statistically significant. This does not mean that training interviewers will have little if no effect. The interviewers in the LFS only differed in terms of the amount of hands-on experience. Training them specifically in terms of understanding the concepts and structure of the classification system itself may prove to be advantageous.

Analysis of the success of trigram-coding by section of economic activity indicates that the list of descriptions is the main factor in determining whether this type of coding returns acceptable codes or not. The importance of the coding file is also emphasised by other authors.

The results show that trigram-coding of economic activity prolongs the interview, mainly because it is an added item in the questionnaire. There is also reason to believe that this type of coding takes somewhat longer to record than the average interview item. Our crude measurement did not, however, enable us to verify this. As borne out by the study of Roessingh and Bethlehem (1993), failures to assign a code result in considerably prolonging the interview. In this there was a marked difference between the experienced and inexperienced interviewers, clearly showing that better training may lead to better results.

In this study the interactive coding performed only marginally better than manual coding in terms of costs, as measured by time. Different pay structures may change the odds more distinctly in favour of trigram-coding. We did not take into account in this study the costs of developing the trigram module, nor the cost of training the interviewers. It is, however, not clear whether the training cost should be included. If there is a reason to train the interviewers to record better descriptive data for trigram-coding,

there is certainly a reason to do so in order to improve the quality of the data for manual coding. Furthermore, neither the investment in developing a computer-assisted manual coding facility in MS Access nor the investment in training the expert coder were taken into account.

6. Conclusions

The results of this study show that as a means of coding economic activity during the interview, the trigram-coding facility in BLAISE III is a viable instrument. The total success rate, prior to editing, is at acceptable levels and the cost of trigram-coding is less than the cost of exclusively manual coding. The differences in costs are, however, not great. The study indicates that the performance can be considerably improved, both in terms of the success rate and in terms of costs.

First, the list of coding descriptions can be refined in order to increase the likelihood of hits, especially within certain sections of economic activity.

Second, training of interviewers is important. The training should both consider interview techniques, such as knowing when there is time to quit searching in the trigram dictionary, and the concepts and structure of the classification system. The latter, however, was not tested in this study.

Finally, interpretation of these results as valid for all circumstances of trigram-coding should be made with caution. Classification systems can be vastly different as the underlying structures may be different in how they lend themselves to codification. These differences can also be across language and cultural boundaries. Interactive coding of a concept such as economic activity may thus not be feasible in all languages.

References

Bushnell, D. (1995). Computer-assisted Occupation Coding. In Essays on Blaise 1995, Proceedings of the Third International Blaise Users' Conference, V. Kuusela (ed.), pp. 23-33, Helsinki, Statistics Finland.

ILO (1990) *Surveys of economically active population, employment, unemployment and underemployment.* An ILO manual on concepts and methods. Geneva.

Institute of Economic Studies (1997). Menntun, mannauður og framleiðni [Education, human capital and productivity] (in Icelandic). Hagfræðistofnun Háskóla Íslands, Report no. C97:02, Reykjavik.

Lyberg, L. and Dean, P. (1992). Automated Coding of Survey Responses: An International Review. R & D report 1992:2, Green series, Statistics Sweden.

Martin, B. (1993). The Blaise Coding Facilities - a closer look. Paper presented at the Second International Blaise Users Conference, London, October 1993.

Roessingh, M. and Bethlehem, J. (1993). Trigram-coding in the Family Expenditure Survey of the CBS. In Essays on Blaise 1993, Proceedings of the Second International Blaise Users Conference, pp. 118-132, London, Office of Population Censuses and Surveys.