REPORT ON SPOKEN LANGUAGE RECOGNITION YEAR 2000 CENSUS QUESTIONNAIRE

Lawrence A. Malakhoff, Martin V. Appel, Bureau of the Census, Washington, DC 20233
Ronald Cole, Center for Spoken Language Understanding, Oregon Graduate Institute
of Science &Technology, Beaverton, Oregon
Lawrence A. Malakhoff, U.S. Bureau of the Census

1.0 BACKGROUND

The U.S. Bureau of the Census is a large, general-purpose, statistical agency. It conducts the Census of Population and Housing (Decennial Census) in years ending in 0, the Economic and Agricultural Censuses in years ending in 2 and 7, and hundreds of establishment and household surveys on a biannual, annual, monthly, or weekly schedule. The Decennial Census is by far the Bureau's largest and most complex data collection activity. Until quite recently, almost all the Bureau's data collection activities used paper questionnaires. This is changing.

2.0 Technology Definition

An Automated Spoken-language Questionnaire (ASQ) is a technology which allows a respondent, speaking into a telephone, to reply to computer generated prompts. The ASQ system functions as an interviewer. The system answers a call, prompts the respondent, recognizes the respondent's spoken replies, and stores the collected information. Data items, if not properly recognized, can be recorded for future data keying. Respondents have the ability to "bail-out" to a human operator if they desire.

3.0 Motivation

While it is very likely that the primary data collection vehicle will still be the paper questionnaire, during certain periods of the data collection process other methodologies may be advantageous. Potentially, the ASQ offers a number of advantages over a written questionnaire: 1) a respondent need not be literate; 2) responses can be interpreted by the system and repaired by the respondent; and 3) upon completion of the ASQ, the data is available for processing. For example, when a respondent calls in to request a questionnaire or Census calls the respondent to request data, an ASQ would be superior to waiting for the mailed questionnaire or having the telephone interviewer key the data.

By its very nature, the Bureau's traditional short form census questionnaire favors the introduction of telephone technology. It contains only a few questions which, if the dialogue is structured properly, can be answered with a constrained vocabulary. This technology would maintain a natural interaction with the respondent, while eliciting answers equivalent to those that would have been entered on a paper questionnaire.

4.0 Evaluation Process

The evaluation of the ASQ data collection methodology followed the three step process recommended by the Census Bureau: 1) initial technology assessment; 2) small scale feasibility testing; and 3) operational testing.

The first step, an **initial technical assessment** (ITA), answered questions covering such topics as: range of potential survey and census uses; difficulty of application setup; costs of initial investment; user training required; user acceptance; effects on coverage, response rates, estimates, and timeliness. In 1993, the ITA evaluation committee recommended a feasibility study and a research plan for a short answer survey. [1]

5.0 Research Plan

To explore the potential of ASQ, the Census Bureau, through the Office of Naval Research (ONR), commissioned the Oregon Graduate Institute Center for Spoken Language Understanding (OGI) to build a proof-of-concept (POC) ASQ system that modeled a subset of the Decennial Census short form questionnaire.

A three tier Research Plan was developed to complete the three step evaluation process. A production prototype system was developed to acquire information from callers in English [2]. Tier 1 consisted of the following: 1) determining the dialogue structures, 2) collecting and transcribing the speech data needed to design and train the prototype systems, 3) developing semantic and dialogue models, and 4) developing a POC-ASQ system.

Tier 2 consisted of a **small scale feasibility test** of the delivered POC-ASQ system by the Bureau's Center for Survey Methods Research (CSMR) to evaluate it in terms of the Census Bureau's goals.

Tier 3 research was the **operational test** of ASQ. It involved engineering, deploying, maintaining, and evaluating an ASQ production prototype used in the 1995 Test Census conducted in Oakland, California; Paterson-Clifton, New Jersey; and six parishes in northwest Louisiana.

5.1 Tier 1 Research

The specific objective of Tier 1 research was to develop a POC-ASQ system for use in the 1995 Test Census. The dialogue of the questionnaire was determined by the analysis of caller's responses for

three different protocols. The most effective phrasing for each question was retained, others were reworded until a satisfactory dialogue emerged. Speech data were collected nationwide to design and train the prototype system [3]. The POC-ASQ system collected the following information from callers about themselves:

1. Last name, first name, middle initial
2. Gender
3. Date of Birth
4. Marital status
5. Spanish/Hispanic origin
6. Race
7. Home Telephone Number with Area code

5.2 Tier 2 Research

Tier 2 research was a behavioral study of the POC-ASQ. In June and July 1994, the Center for Survey Methods Research (CSMR) undertook the task of evaluating respondent's attitudes towards the ASQ [4]. A marketing research firm recruited 20 people for a small-scale test of the POC-ASQ system. These subjects were selected by income, race, age, and gender to get a representative sample of the U.S. population.

CSMR collected suggestions for improving the ASQ from their test subjects. These suggestions focussed on three areas:

Correcting Errors

Errors should be corrected as they occur rather than waiting until the end of ASQ. subjects said it was difficult to remember all errors.

Pace

The pace of the interview needed to be sped up for most participants.

Question wording

Recommendations for rephrasing questions emerged for the Hispanic origin and race questions.

Besides a test of the POC-ASQ, CSMR conducted similar research with a live interviewer. The results of their study showed that the POC-ASQ performed nearly as well as the live interviewer for most characteristics. It should be cautioned that since the sample size was so small, broad conclusions should not be made about caller preferences. However, the trend from this study shows that no feature of the POC-ASQ surfaced that would cause a respondent to be against using this technology.

5.3 ASQ Goals

•Operator Bailout

Add the capability to hand a call to an operator. The system should transfer the caller to a human operator when it judges that it cannot handle the call, or the caller requests to speak to an operator.

• Include Coverage questions

These questions usually require yes/no answers. For example, the ASQ could ask if other members of a household are temporarily living elsewhere.

• Include all household members

The ASQ asks for personal data about all household members and their relationship to the caller.

• Detection of non-responsive utterances

The system needs to determine if a response is a word or some other sound before asking the next question.

• Improve Breakdown Repair

This item deals with how the system manages unrecognized responses, or responses with low confidence.

• Operator Graphical User Interface (GUI)

This GUI allows an operator to review or edit calls made to the ASQ or assist callers who bailed out of the system.

• Barge-in Capability

This capability allows a caller to respond to a question before it has finished playing.

• Better Recognition of Spelled Names

The caller spells his or her name. Since some letters sound alike, the result is compared to a database of names to capture a name correctly [5].

• Robust Recognition in the presence of Noise

Respondents may be calling from locations where there may be background or line noise present.

6.0 ASQ Evaluation

The completed ASQ was not available for remote testing from OGI's laboratory in early January, 1995. According to OGI, the delay in testing was due to the task of constructing the speech recognizers for words describing the relationship between the caller and other household members. This delay in testing the production system prompted the Census Bureau to send two people to the OGI lab to conduct an assessment of the system's working components on January 30-31.

6.1 ASQ Laboratory Evaluation

The first priority in this assessment was to verify that the newest version of the protocol was implemented as specified, to include coverage questions and collect information about all household members. The system was called numerous times to verify that proper branching to the next question for recognized responses occurred. At that time, there were recognition

problems with the id number, race names, and words describing the relationship between the caller and household members. The instruction to bail out to a human operator did not function for recorded responses. This instruction was inserted after the instructions to record household members' names.

Other ASQ goals assessed were the barge-in capability, background noise handling, and recognition of spelled names. The barge-in application program interface between OGI's software on the digital signal processing (DSP) chip on the LINKON telephone board and LINKON's DSP software was still being developed. Background noise handling was said to have improved, but how much was not measured. Files of male and female first names, and last names, derived from the 1990 Census data and 1993 IRS 1040 returns, were merged with OGI's name files. These name files were used to improve the recognition of spelled names.

Operator bailout functioned from a caller's perspective; the system switched a call promptly to an operator. The operator screens were not fully developed at this time. The main problem was the difficulty going from one question field to the next. The screens needed different colors to improve the contrast of the text and larger fonts to make the text more readable.

Other questions concerned how calls were managed by the operator consoles. Since calls would not flow in continuously, an audio signal was proposed to alert an operator of an incoming call. A different audio signal was used to alert an operator that he or she was receiving a bailout. An operator who is idle or editing a call could then click on the "Accept" button and take the call. A caller who bails out would have all their previous answers available for the operator to review; question screens would pop up as the operator completes the interview. The call editing screens would enable an operator to retrieve a past call to edit while calls were streaming in or at the end of the processing day. These procedures were implemented for the production ASQ.

System response time is the time it takes for the system to recognize when a person has finished speaking, recognize a response, and return with the next question. The CSMR study stated that the pace of the interview was one of three main points that needed to be improved for the production ASQ. With a shorter system response time, the pace of the interview would be increased. Response time was improved from a four to five second delay to a two to three second delay over the POC-ASQ. Questions were also posed about how system load would affect response time. System response time began to degrade when four or more calls were made simultaneously during this evaluation.

ASQ System Configuration

The Census ASQ system was distributed over several platforms. It used a digital (T-1) telephone line, providing 20 inbound and 4 outbound operator channels. Calls were forwarded to the Jeffersonville, IN FTS2000 system by CATI interviewers in Tucson. Calls could also be sent to the ASQ via a touchtone option on a menu offered on a 1-800 number. The T-1 line was connected through a channel bank to three LINKON voice boards in a PC-class computer running the Solaris operating system. When recognition was required, processing was sent over a LAN to one of 6 DEC Alpha computers. Each Alpha was designated for a specific recognition task to shorten processing time. Four Alphas also served as operator consoles for bailout and editing.

6.2 ASQ Production Evaluation

Installation of the production ASQ system began on February 9, but was not completed until March 7 because of an incompatibility problem between the OGI software and the LINKON telephone board. The ASQ was offered on the 800 menu that day.

The system received its first call on March 14. Individuals who reported their data to the Tucson facility via CATI were also given an option to use the ASQ. The lack of activity since March 7 was because the wrong check digit algorithm was used for validating a caller's id on the Tucson CATI system. This error caused 560 calls to be flagged as invalid. Between March 15 and April 5, the system received 16 more calls. In total, 108 calls were eligible for the ASQ between March 4 and April 10. Eligible callers had to be reporting for the short form, spoke English, and wished to use ASQ. About 16% of those eligible chose the ASQ.

Public Call-in Period

The ASQ system was operating most of the time the public was allowed to call in. The system crashed three times; once, due to hardware failure, and two other times because of computer memory problems. On these occasions, Jeffersonville personnel called Tucson and informed them not to transfer calls. No respondents were disconnected by the ASQ when the system crashed. However, recovery procedures given to Jeffersonville personnel to restart all the recognition processes did not work. OGI programmers had to be contacted each time the system crashed. Operator bail-out worked and the bail-out and editing screens were being used by the Jeffersonville operators without any problems.

Load Test

The trickle of calls that came into the system

during the public call-in period was not enough to assess the capability of the ASQ. A load test was arranged using headquarters personnel to call the system on April 11 & 12. Callers placed a heavy load on the system, sometimes busying out all the available telephone lines. More than eight hundred calls were attempted to the ASQ on these days, 359 calls were logged into the system. Initially, the system was configured to accept twenty simultaneous calls. The system was reprogrammed for eight lines after it was discovered that the PC with the LINKON telephone board was having memory problems and either disconnecting callers or taking more than five seconds to play the next prompt. Memory problems also occurred with the Avanti DEC Alphas. These machines could not process voice recognition computation and do call editing at the same time. Therefore, operators had to edit the calls at the end of the day.

Public Call-in Period Call Volume

We speculate that there are three possible reasons for the low call volume for the ASQ in this time period. Analysis of the call logs revealed that 3 calls were forwarded to the ASQ from the Tucson CATI site, 14 calls came through the 1-800 menu. The first two reasons are about the wording of the mailed out questionnaire and the 800 menu design. The wording of the invitation to use the 800 service line on the mailed out questionnaire did not mention the ASQ option. Further, the wording gave the impression that the respondent would talk to a human. When respondents called, they had to make several menu selections before being asked by a human operator if they wanted to use the ASQ. Respondent comments recorded by our operators indicated strong frustration with this process. Potentially, more calls might have been forwarded to the ASQ if the option had been announced at the beginning of the menu, and the printed invitation made clear what a respondent was going to experience.

The third reason call volume was low was that the ASQ software and hardware was unstable. On two occasions the system crashed and was down almost two full days. Given that 16% of callers responded to the ASQ when offered, the call volume during the test was about 1 call every two days. Potentially, one or two calls may have been missed due to the ASQ software and hardware problems. Seventeen valid calls were made to the system [6]. Eight of these callers hung up before completing the questionnaire, seven bailed out to an operator, and two answered all the questions.

7.0 ASQ Prototype Performance

A load test of the ASQ was performed April 11th & 12th. The system was running for 19 hours and received 359 calls, although more than eight hundred were attempted. Table 1 describes the effectiveness of the system to obtain the correct information from the caller. The total number of responses for each type of prompt is listed, along with the percentage of the time the respondent was reprompted. Column three contains the number of responses that contained the desired information, a response with a key word, followed by a count of responses without the key word. The last column lists responses that were recognized, and those meaningful to a Census operator who listened to, and transcribed spoken data.

TABLE 1
QUESTIONNAIRE EFFECTIVENESS

Question	Responses	Reprompt (%)	Key Word Obtained	No Key Word	Total Usable
yes_no	2520	10	2483	37	2491
firstname	289	20	244	45	275
lastname	305	30	209	96	268
sex	277	10	274	3	274
day	328	40	317	11	319
month	297	30	278	19	285
year	246	0	237	19	239
relation	82	20	80	2	81
association	34	10	32	2	33
origin	37	20	36	1	36
race1	239	10	229	10	229
race2	31	0	28	3	28
race3	21	0	21	0	21
race4	13	0	13	0	13

Table 2 details recognition performance for the two day load test.

TABLE 2
RECOGNITION PERFORMANCE

Question	Recognition (%)	Low Confidence	Other Errors	2nd Pass Accuracy (%)
yes_no	98.75	34	1	99.81
firstname	81.09	50	34	92.53
lastname	64.55	87	68	86.50
sex	98.54	30	3	99.63
day	72.73	120	65	91.33
month	93.29	15	2	100.00
year	68.62	0	0	68.62
relation	69.14	22	9	77.78
association	63.64	0	0	63.64
origin	72.22	0	0	72.22
race1	96.94	14	3	98.23
race2	100.00	0	0	100.00
race3	78.43	0	0	78.43
race4	100.00	0	0	100.00

8.0 OGI Field Test

OGI performed their own independent test of the ASQ with 133 respondents. They revised and shortened the protocol used in the 1995 Census test. Their results are not reported here because their data is not directly comparable to Census test data.

9.0 Conclusions

We recognize that the implementation of the production prototype ASQ was less than optimal. We would have preferred a more systematic research plan; one with a treatment group that is offered the ASQ option, and a control group that is not. At this point in time, this was not possible, given all the other research that is being conducted within the 1995 Census Test.

The caller comments from the headquarters calls are still being transcribed. The problems that occurred were mostly engineering problems involved with scaling up a laboratory product to a production test. The telephone communication problems were secondary to computer memory problems. As the speed of the disk controller and bus of PCs increases, memory and response time problems should be reduced. Presently, prototype recognizers are being tested that have real-time system response and barge-in capability.

The decision has been made not to use voice recognition for the year 2000 census. Time and resources are scarce. The effort to correct and test the problems that occurred during the test could seriously affect higher priority technologies.

The computer hardware and software from the test have been moved to headquarters. Once a thorough analysis of the data has been completed, a decision about voice recognition at the Census Bureau will be made.

References

Appel, M.V., (1993). "Voice Recognition Entry, Initial Technical Assessment." CASIC Committee on Technology Testing, 1993.

Appel, M.V., Cole, R. (1994). "Spoken Language Recognition of the Year 2000 Census Questionnaire, A Feasibility Test." Proceedings of the American Association for Public Opinion Research, 1994.

Cole, R., Novick, D.G., Fanty, M., Vermeulen, P., Sutton, S. (1994). "A Prototype Automated Spoken-Language Questionnaire for the Year 2000 Census." *Speech Communication, Special Issue*, December, 1994.

Jenkins, C.R., Appel, M.V. (1995). "Respondents Attitudes Towards a U.S. Census Voice-Recognition Questionnaire." International Field Directors and Technologies Conference, 1995.

Cole, R., Fanty, M., Roginski, K. (1992). "Recognizing Spelled Names with Telephone Speech." Proceedings of Speech Tech/Voice Systems Worldwide 1992.

Cole, R. A., Novick, D. G., Vermeulen, P. J. E., Sutton, S., Fanty, M., Wessels, L. F. A., de Villiers, J. H., Schalkwyk, J., Hansen, B., Burnett, D., (1995). "Experiments With A Spoken Dialogue System For Taking The U.S. Census," under review, submitted for publication in the *"International Journal of Human Computer Interaction."*