INFORMATION STRUCTURES IN SURVEY INSTRUMENTS

Naomi Sager, Paul Mattick, Jr, Carol Friedman, Emile C. Chi, Courant Institute of Mathematical Sciences, New York University
Naomi Sager, 251 Mercer St., New York, NY 10012

ABSTRACT

Advances in mathematical linguistics have shown that a study of patterns of word combination in scientific writing yields characteristic word classes and proposition types that carry the information of the science. Computer programs based on these methods have been used to map free text input from technical documents into a relational database whose tables reflect the informational structures in the textual material. This paper reports initial results of applying these methods of text analysis and database design to survey instruments. Queries directed to a pilot database of analyzed survey questions on employment, income and program participation were able to retrieve questions from different instruments in terms of generic informational categories. Some queries that have been executed on the pilot database (where R stands for the respondent) are: (1) In this survey what is the range of relations to a job that R might have; (2) Find all questions in which R has some condition related to employment which qualifies R for income or program participation; (3) Generate a KWIC index of questions on employment. The analysis also provides an objective basis for determining the informational complexity of survey questions.

METHODS

The work reported here has as its long range goal the development of objective frameworks for information in the social sciences. The methods for discovering these frameworks are based in a principled way on a computable analysis of the language of the science; hence the documents of a given field can be projected by computer programs onto the informational frameworks (e.g. databases or special formats) for that field. Projecting a document onto such frameworks does not alter the information. The frameworks specify the location of each kind of information, so that procedures that operate on the analyzed documents can find each kind of information if it is present. This makes possible a wide range of computer applications for storage, retrieval, and comparison of portions of documents in respect to their information content and to such features as their wording, their grammatical form, and the amount of information in the individual sentences.

With regard to the methods of analysis, the underlying observation is that the verbal material in a specialized area naturally falls into characteristic statement types that consist of the special vocabulary classes of the area as they occur in particular syntactic combinations. Once determined for an area, these statement types (and their characteristic combinations in turn) constitute the underlying informational structures of textual material in that field. They provide generic informational categories and relations in terms of which the content of different documents in the field can be compared and diverse high level informational tasks programmed to operate on the linguistically structured representation of the texts.

The methods have previously been applied to biomedical fields, chiefly laboratory sciences and medical records, to obtain canonical information formulas, or "information formats", that organize the information in the textual material of the subject area [1,2,3]. The feasibility study reported here investigated whether the same methods that had proved successful in natural science texts could be applied to social science materials, in particular to survey instruments, where codification of verbally-obtained data is already well-developed and where questions regarding the comparability of data obtained by linguistically similar questions have been raised [4].

APPLICATION TO SURVEY INSTRUMENTS

A feasibility study in the applicability of sublanguage methodology to survey instruments was made. The data of the study consisted of portions of

Survey of Income and Program Participation (SIPP 5100 - 1985 Wave 1)
Study of Family Economics (PSID-SRC 1985)
Survey of Work Experience, of Mature Men (NLS LGT 1121 1983).

The methods of linguistic analysis referred to above (see also [5]) were applied to the sample data set to find the classes of words and patterns of word class cooccurrence that appear regularly in the survey questions in the data set. Each elementary word class pattern, or "statement type" corresponds to one type of information that appears in the textual material undergoing analysis. The application of these methods of analysis to the pilot data set of survey questions yielded three main statement types, which appear in the survey questions in various combinations, joined by linguistic connectives to form the more complex questions.

A pilot relational database was then designed whose structures correspond to the statement types identified by the sublanguage analysis, and an implementation of the database using dBASE III on a personal computer was undertaken. The text of the questions in the data set was keyed in as well as the structured form of the questions, and a set of 20 sample retrieval queries (Table 1) was programmed so that an online demonstration of the database could be performed. Q1 in Table 1, for example, asks: IN THIS SURVEY WHAT IS THE RANGE OF RELATIONS TO A JOB THAT R [the respondent] MIGHT HAVE? Note that this high-level query could not be replaced by a key-word search.

In the pilot database the three main statement types appear as three database tables, illustrated in Table 2: the EMPLOY table (type and duration of employment), the INCOME/PROGRAMS table (income and program participation), and the STATUS table (conditions such as health, which might qualify employment or program participation). To illustrate, the top section of Table 2 shows an instance of the EMPLOY statement type as it appears as a row of the EMPLOY database table. It corresponds to the SIPP 2.d question written just below it in the table, namely:

> WHAT WAS THE MAIN REASON R [the respondent] COULD NOT TAKE A JOB DURING THOSE WEEKS?

The table has fields for the characteristic subject, verb, and object of an EMPLOY unit; the SUBject here is the respondent R, the object of the verb, labeled JOB, is the word job in this case, and the verb VJOB that relates the subject to the job-word object here is *could not take*. There is also a DURation entry: *during those weeks*, and a WHY or *reason* entry, in this case the item being questioned: *What main reason*. Each table contains from 10-20 fields which include in addition to the "core" elements of the statement type (the named columns illustrated in Table 2), a field that indicates which element is being questioned, and fields for modifiers.

Survey questions may contain more than one informational unit, where each instance of a statement type constitutes one informational unit. A CONNECTIVE table in the database records for each complex question the interrelation of its informational components. For example, Figure 1 shows the

TABLE 1: SAMPLE QUERIES TO SURVEY DATABASE

Q1 In this survey what is the range of relations to a job that R might have?

Q2 List all questions where the type of relation which R has to employment is similar.

Q3 What sources of income or program payments are asked about in the surveys in the database?

Q4 List all questions that relate R's non-wage income to R's qualifying condition. See handout for connective tree structure.

Q5 List all questions on R's personal assets.

Q6 Display all questions on R's income and program participation.

Q7a Display all questions on R's income from Federal government sources.
Q7b Display all questions on R's income from non-government sources.

Q7c Display all questions on non-monetary assistance programs.
Q7d Display all questions on non-governmental support payments.

Q8 Generate a KWIC index of questions on:
(1) income or program participation
(2) health plans
(3) personal assets

Q9 Find questions in which R has some condition related to employment

Q10 Find questions in which R has some condition related to employment which qualifies R for income or program participation.

Q11 Display questions re employment with the same job verb and job word.

Q12 Display questions with a similar element re employment.
Q12a Same, formatted for printed output.

Q13 List all questions which refer to R's coverage by health plans.

Q14 List all questions about R's coverage by MEDICAID.

Q15 List all questions about R's coverage by MEDICARE.

Q16 List all questions which refer to R's coverage by health plans other than Medicaid or Medicare.

Q17 List all questions on receiving Social Security.

Q18 List all questions on receiving welfare.

Q19 List all questions about the personal status of R.

TABLE 2: **BASIC QUESTION TYPES**
(Examples from SIPP 5100 data set)

1. EMPLOY

SUB	VJOB	JOB	DUR	PER	PAY	WHY
respondent	job-verb	job	duration of job	period surveyed	pay	reason
R	*could not take*	*job*	*during those weeks*			*what main reason*

for SIPP 2.d.: *What was the main reason R could not take a job during those weeks?*

2. INCOME/PROGRAM

SUB	VREC	INC	TYPE	SOURCE	DUR	PER
respondent	recipience-verb	income	typeof income	source of income	duration of recipiency	period of survey
R	*receive*	*any money*	*from worker's compensation*		*during this 4-month period*	

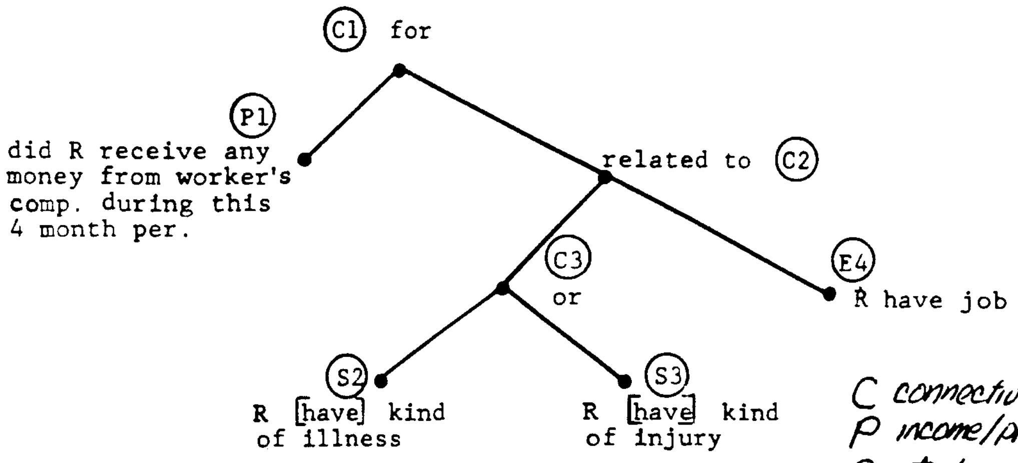
for SIPP 10: *During this 4-month period did R receive any money from worker's compensation* for any kind of job-related illness or injury?

3. STATUS

SUB	VSTAT	STAT
respondent	status verb	status
R	*have*	*any kind of illness*
R	*have*	*any kind of injury*

for SIPP 10: During this 4-month period did R receive any money from worker's compensation for *any kind of* job-related *illness or injury?*

FIGURE 1: CONNECTIVE Structure for SIPP 10


C1 for
P1
did R receive any money from worker's comp. during this 4 month per.
related to C2
C3 or
E4 R have job
S2 R [have] kind of illness
S3 R [have] kind of injury

C connective
P income/program
S status
E employ
P,S,E numbered consecutively

connective structure for Question 10 of the SIPP 5100 data set: DURING THIS 4-MONTH PERIOD DID R RECEIVE ANY MONEY FROM WORKMEN'S COMPENSATION FOR ANY KIND OF JOB-RELATED ILLNESS OR INJURY? The question contains four instances of statement types, that is, four units of information are combined in this question. The main clause, labeled Pl in the diagram, is an instance of an INCOME/PROGRAM statement type in question form: DID R RECEIVE ANY MONEY FROM WORKMEN'S COMPENSATION DURING THIS 4-MONTH PERIOD? P1 is qualified by an EMPLOY unit E4, linked to 2 alternative STATUS units S2 and S3 via the words *related to*. That is: Did R have any kind of illness? Did R have any kind of injury? And were either of these related to R's having a job? All of these are then related to the PROGRAM PARTICIPATION unit P1 by the connective *for*.

The point of identifying the component statement types and their interconnections within a survey question is that one is then able to retrieve all questions that utilize a particular type of information, and also those that utilize these types in particular combinations, even if they are expressed in different ways or appear in combination with otner elements. Such a database could help researchers who are looking for a particular subpopulation; for example, one defined by personal characteristics and status variables. An investigator might want to find people who have experienced a given type of event, such as "received welfare", "married", especially in combination with other variables, as is expressed in the demonstration database by links among particular rows in the relational tables. E.g. *Did R receive sick pay when he was ill?* would be expressed by instances of the INCOME/PROGRAMS and STATUS relations in the database, where the STATUS relation contains an *illness* word, and the two instances are related in the CONNECTIVE table (i.e. occur under a connective). This question and the SIPP 10 question would both be retrieved by a query to the database asking for all questions that relate a respondent's receipt of income or program participation to illness, because both questions contain an underlying informational structure that is realized in the database structure. Again, queries employing such generalized informational categories are not achievable by keyword search.

Other informationally complex retrievals from the sample database have been programmed and illustrate the potential for survey methods research of having an informationally structured database of survey instruments. The next step of analysis is to include multiple choice answers and the text of answers to open-ended questions. Such extensions of the research are planned.

REFERENCES

[1] Z. Harris, M. Gottfried, T. Ryckman, P. Mattick, jr, A. Daladier, T.N. Harris and S. Harris, *The Form of Information in Science: A Test-Case in Immunology*, Boston Studies in the Philosophy of Science, Reidel, Dordrecht 1987.

[2] Sager, N., *Natural Language Information Processing: A Computer Grammar of English and Its Applications.* Addison-Wesley, Reading, MA, 1981.

[3] Sager, N., Friedman, C., Lyman, M.S., MD, and members of the Linguistic String Project. *Medical Language Processing: Computer Management of Narrative Data.* Addison-Wesley, Reading, MA, 1987.

[4] Schuman, H., and Presser, S., *Questions and Answers in Attitude Surveys: Experiments in Question Form, Wording, and Context.* Academic Press, 1981.

[5] Z. Harris, *Language and Information*, The Bampton Lectures in America for 1986, at Columbia University, Columbia University Press, New York 1987.