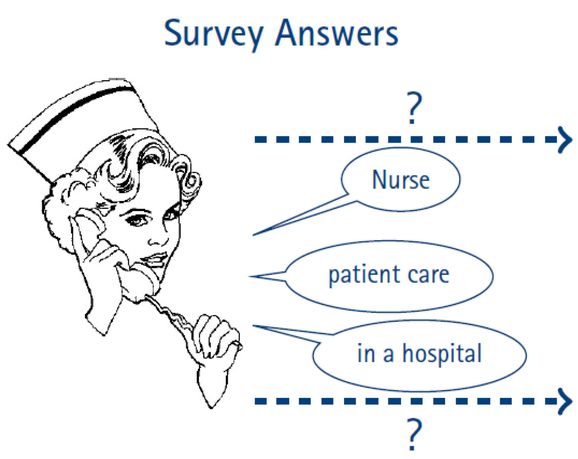
A2.7 New Methods for Job and Occupation Classification

Directors/ Frauke Kreuter
Researchers/ Malte Schierholz
Funding/ DFG

The Coding Problem



Classification

KldB 2010	Job Title
81302	Nurse
81313	Graduated Nurse
81323	Pediatric Nurse
⋮	⋮
81102	Doctor's Assistant
82102	Geriatric Nurse

- Most surveys ask for occupation with open-ended questions. The answers are coded afterwards into a classification with hundreds of categories and thousands of jobs.
- In this example, category 81302 (Nurse) appears to be most plausible but other categories are possible as well. How can we increase data quality?
- Coding is mostly manual work and therefore expensive.

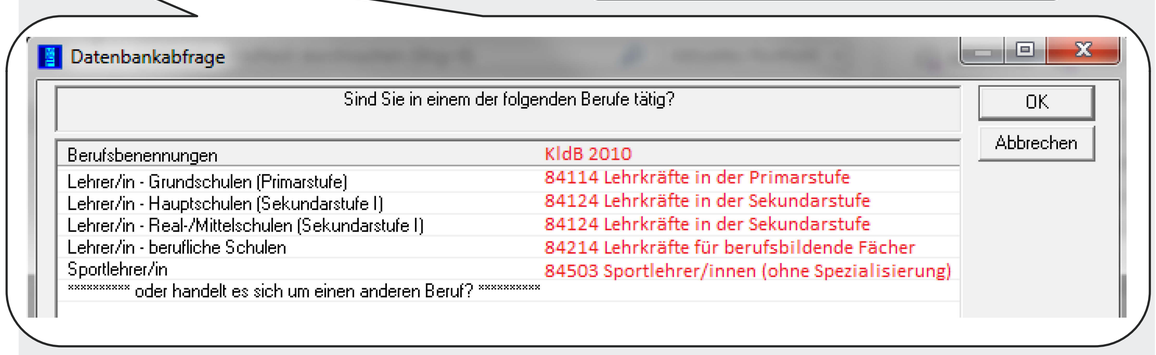
The Solution

"Please tell me which occupational activity you pursue."

"Vice principal and teacher"

"Please describe this occupational activity precisely."

"German, Maths, and Arts"



When the last option is selected: manual coding as hitherto

Objectives

This project develops a new instrument for the measurement of occupation. The key idea is to utilize supervised learning algorithms to predict candidate job categories. These suggestions are presented to the respondent, who can in turn choose the most adequate occupation.

We expect to improve current practice with respect to the following:

- Reducing the number of observations that have to be coded manually
- Increasing data quality for respondents whose first response is ambiguous
- Saving valuable interview time by automatic question selection

Methods & Data

Automatic Job Prediction

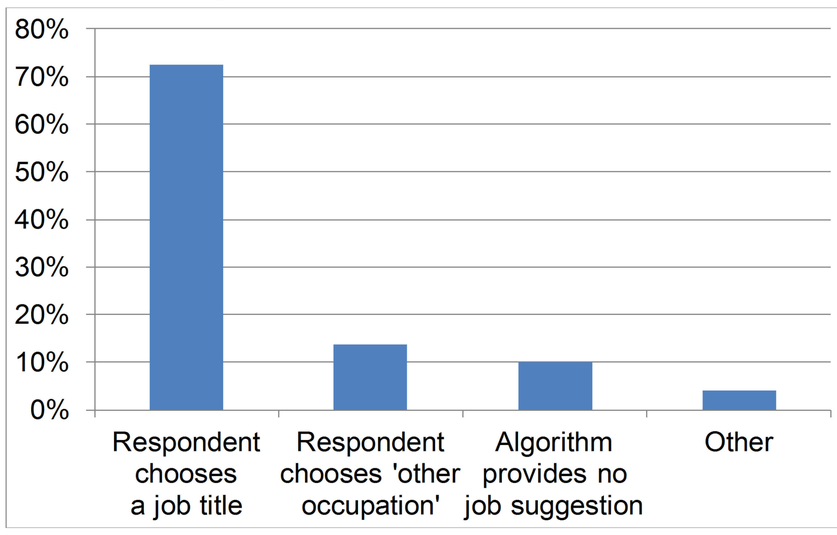
- Combination of supervised learning and rule-based prediction
- Learning system that will improve when more coded answers are fed into the system
- The five most probable job titles are suggested to the respondent

Live Test

- Tested in a phone survey operated by TNS Infratest Sozialforschung
- 1064 employed respondents
- Two professional coders for independent comparisons
- Two student assistants assessed the quality of all three coding procedures

Results

1. Productivity



- Interview coding successful for 72.4% of the respondents
- 23.6% residual cases need to be coded manually …
  - … because the algorithm provides no job suggestions or
  - … because the respondent chooses 'other occupation'
- Special experimental conditions for 4% of the respondents

2. Findings

Quality

- Agreement rates between the interview coding procedure and professional coding procedures are approximately equal.
- Student assistants find that interview coding does slightly worse than the first professional coder but is comparable to the second one.
- Systematic errors happen in few cases when respondents select overly general job titles.

Interview duration

- The proposed adaptive question could replace the additional open-ended question, saving a few seconds.
- Only for respondents who choose 'other occupation' (13,6%), this 2nd question is still required for manual coding, increasing interview duration by 37 seconds.

3. Conclusion

Coding during the interview is technically feasible but complex.

- Overall results are highly promising but several deficiencies were identified as well.

Future Plans

A renewed test is envisaged. The following improvements are planned:

- Interview duration could be reduced if respondents who would probably choose 'other occupation' don't get job suggestions.
- More training data and a better prediction algorithm can improve the whole system.
- An auxiliary classification is needed …
  - ... with precise occupational descriptions to avoid that respondents select the wrong category in the first place.
  - ... which allows unique mappings into both official classifications, KldB 2010 and ISCO-08, for simultaneous coding.
- More attention needs to be paid to the implementation of standardized interviewing.
- Allow coding into more than one category?