

METHODS BRIEFING 26

The Development of Methods for Complex Coding

Peter Elias and Rob Jones

Institute for Employment Research,
University of Warwick

September 2003 - August 2005

Cascot: Computer Assisted Structured Coding of Text

<http://www2.warwick.ac.uk/fac/soc/ier/publications/software/cascot/>

A series of methods briefings from projects funded by ESRC as part of the Research Methods Programme.

The Programme aims to develop qualitative and quantitative methods within the context of substantive research. It also aims to encourage effective dissemination of good practice.

Further copies are available from:
CCSR
2nd Floor
Crawford House
Manchester
M13 9PL

0161 275 4891

www.ccsr.ac.uk/methods/

Cascot is designed to assign a code to a piece of text. In the case of the Standard Occupation Classification (SOC) this piece of text is typically a job title. For the Standard Industrial Classification (SIC) the text is a description of the main product or services provided by an employing establishment. The quality of coding performed by Cascot depends on the quality of the input text.

The software is capable of occupational coding and industrial coding to the UK standards developed by the UK Office for National Statistics. These are the Standard Occupational Classification (SOC) and the Standard Industrial Classification (SIC). Cascot currently supports SOC 2000, SIC 92, SOC 90, SIC 80, and SIC 2003. (For more information on these classifications go to the ONS website at: [//www.statistics.gov.uk/methods_quality/ns_sec/soc2000.asp](http://www.statistics.gov.uk/methods_quality/ns_sec/soc2000.asp))

Occupational coding and industrial coding arise in a number of situations. Examples include job titles which may be entered as free text on surveys or administrative databases. A job title is indicative of the kind of work people do or would like to do, or the sorts of jobs in which employers want people to work. Information like this is collected routinely in a wide range of settings such as job vacancy advertising, careers guidance or official statistical enquiries. Coding is the process of categorising the huge range of all possible answers to a pre defined set of categories (each category having a unique code).

Ideally the text should contain sufficient information to distinguish it from alternative text descriptions which may be coded to other categories within the classification, but it should not contain superfluous words. This ideal will not always be met but Cascot has been designed to perform a complicated

analysis of the words in the text, comparing them to the words in the classification, in order to provide a list of recommendations. If the input text is not sufficiently distinctive it may not be the top most recommendation that is the correct code.

When Cascot assigns a code to a piece of text it also calculates a score from 1 to 100 which represents the degree of certainty that the given code is the correct. When Cascot encounters a word or phrases that is descriptive of occupation or industry but lacks sufficient information to distinguish it from other categories (i.e. without any further qualifying terms) Cascot will attempt to suggest a code but the score is limited to below 40 to indicate the uncertainty associated with the suggestion. For example 'Teacher' or 'Engineer' come into this category.

The performance of Cascot has been compared to a selection of high quality manually coded data. The overall results show that, for occupation coding, more than 80% of records receive a score greater than 40 and of these more than 80% are matched to manually coded data. When using Cascot you can expect this level of performance with similar data, but be aware that the performance is dependent on the quality of your input data.

How to obtain CASCOT

Cascot is available online for free. A desktop version of the Cascot software suitable for

processing high volumes of data is also available for purchase. Cascot requires a Java 2 Standard Edition (J2SE) Runtime Environment (JRE) version 1.4 or higher. When you download the free version available at the website shown below, a check will be made to see if your system has the required Java environment. If this check shows that you need to obtain or update your Java environment, you will be directed to the website where you can download Java for free.

A desktop version of Cascot is also available. This has more features than the online version, including facilities to process large files of text automatically and semi-automatically and to provide the coder with additional information which may assist the coding process. The desktop version is supplied with Cascot Editor, a software tool that allows the user to 'open up' the coding tool and modify it for specific purpose, or to import their own classification into the coding tool. The desktop version costs £255.32 (excluding V.A.T.) + £44.68 V.A.T. = £300

To buy a copy or to use the online version for free go to:

<http://www2.warwick.ac.uk/fac/soc/ier/publications/software/cascot/>

Further details are available from:
Peter Elias
Email: Peter.Elias@warwick.ac.uk
Phone Number: +44 (0) 2476 523286