 

**DASISH (Data Service Infrastructure for the Social Sciences and Humanities)**

Report of the CASCOT Training Workshop:

Occupational Coding in Multi-national Surveys

held on 10 to 11 April 2014 at

University of Warwick, Palazzo Pesaro Papafava, Cannaregio 3764, 30121 Venice

*Attached to this brief note on the Workshop are the following Appendices:*

1. *Programme and timetable for the Workshop;*
2. *List of participants with contact details;*
3. *Presentations: E Harrison; P Elias, M Birch; R Ellison(x 2); S Westerman; E Meschi, M Belloni*

**Aim of the workshop**

The aim of this project within DASISH is to develop prototype multilingual software that will code national language text to an international classification. The classification selected for this prototype is the *International Standard Classification of Occupations 2008* (ISCO 08). The prototype of the software, known as CASCOT[[1]](#footnote-1), now exists in a multi-lingual version (CASCOT 5.0) in Dutch, French, German, Slovak, Spanish, Italian and Finnish in addition to English. However, development of the English and Dutch versions revealed that a significant degree of ‘fine-tuning’ of the software is required. This is achieved by using the software and its coding rules to re-code job title text that has already been coded to ISCO 08. A specialised Editor has been developed for this purpose.

This Workshop was organised to demonstrate the use of the Editor to improve the performance of the software in different languages. Having received a copy of the software in advance of the Workshop, participants would receive instruction in the use of the Editor and were invited to continue fine-tuning of the software in their national language.

**DASISH and Workpackage 3 *(E Harrison)***

Eric Harrison is involved in the coordination of the European Social Survey (ESS). Attention was drawn to the high relevance of occupation in the context of social class and to the need for transparent ways of establishing the validity of national occupational coding in cross-national survey work. The DASISH project seeks to improve this via the development of three multi-language tools: a questionnaire design and documentation databank; the translation tool for occupational coding; and a question databank, including every question ever asked in ESS and SHARE.

**Development of CASCOT 5.0 *(M Birch, P Elias)***

This brief presentation gave information relating to the background to, and an outline of tasks associated with the development of CASCOT 5.0 within the DASISH project.

**Demonstration of Performance of CASCOT 5.0 *(R Ellison)***

Following a screenshot demonstration of how CASCOT is used for coding text, a screenshot of each language version was shown. Discussion focused on how the probability score for selecting a code is arrived at, and on index size. The result of testing CASCOT firstly for four countries via the ESS Round 6 data was shown, and secondly a comparison between ESCO coding of 4,800 job titles for seven countries and CASCOT. In both cases, CASCOT was used in automatic mode. The CASCOT Performance Tool was used to provide analyses of the results of these tests.

**CASCOT and its Coding Rules *(R Ellison)***

This presentation showed how the CASCOT Editor is used to fine-tune CASCOT. The Rules within the Editor will vary according to each language. Attention was drawn in discussion to the high degree of complexity involved, to examples of differing language requirements (e.g. for those languages in which compound words are very common), and to the impact on established CASCOT coding rules of future changes in the occupational world.

**CBS and CASCOT – tuning CASCOT for improved performance *(S Westerman)***

A specially adapted version of CASCOT has been in use by the Dutch statistical office (CBS) for some time. The presentation detailed how CASCOT is used and analysed its quality and performance. Examples of the way in which the CASCOT Editor rules have been established were demonstrated.

The issue of maintaining the rules was discussed, along with the associated costs that might be incurred. It was pointed out that CASCOT classification files that are approved by the UK statistical office cannot be overwritten. It is, however, possible for users to create their own classification files for their specific purposes. The practice of establishing ‘gold standard’ coding on datasets of 50-60,000 job titles was raised as the key means of maintaining coding quality. Storing the rules as part of the metadata was also discussed.

**Language groups explore the potential to improve CASCOT in specific languages**

This was the final session of day one of the Workshop.

**Feedback from language groups (chaired by PE)**

*Part-word Replacement –* The possibility of establishing new rules for those languages where compound words are common will be investigated with the CASCOT programmer.

*Word ending –* Because the rules relating to word endings are not numbered, this feature of CASCOT did not work too well. Numbering and grouping of these rules to address, for example male/female, singular/plural alternatives should be established to improve performance.

*Processing (or not) of spacing between words –* This is a difficult issue to resolve. Hyphenation software may be useful, but this would not be feasible within the limited resources of this project.

*Text descriptions on screen –* The possibility of building the facility into CASCOT of showing ISCO 08 text description as a pop-up window would be explored, initially in English only (although ISCO 08 descriptions in French, Spanish and German would be available shortly, and unofficial translations will be available in other languages). It was, however, important that this facility was used only for resolving difficult situations where coding via other information was not possible. This should be a locked file, and guidance given as to how it is to be used.

*Methodological issues –* It was important that there was a constant checking of the coding process to ensure accuracy. Semi-automatic processing, checked against expert coding standards, was advisable. In setting new rules within Editor, it should be noted that a new rule may have only a very small impact on the result, so this is a long process. Equally, there is the issue of unintended consequences of establishing a new rule. Gold standard files are very important for checking CASCOT coding results for validity with the performance tool.

Other topics raised in this session included: the use of multiple input fields, other variables and imputation and their effects on data quality; using CASCOT to code to other classifications; dual coding to the national and international classifications vs. correspondence tables; coding by the interviewer at the time of interview (which is OK if accuracy is not vital).

**Recoding SHARE data for Netherlands using CASCOT *(E Meschi, M Belloni)***

This presentation analysed the results of an exercise carried out to compare manual vs. semi-automatic coding of SHARE data using CASCOT.

**Final Comments**

It was agreed that updated classification files would be sent to the IER team, with ideas for improving the software. Developing Czech and Portuguese versions was discussed. Work on this would, however, depend upon the availability of the programmer and the resources that could be devoted to this additional task.

In response to a query from David Hunter, it was confirmed that CASCOT software could be made available free of charge to developing countries.

Clarification is needed from the programmer regarding a detailed description of how the rules are applied.

Future activities within DASISH would include a workshop bringing together the various software packages that have been developed during the project. This was likely to take place during the autumn 2014.

There was a need to bring together NSIs and the academic community to promote gold standard files. DG Research and Innovation infrastructure project may be an appropriate route within the EU.

An interesting project would be to try to bring together CASCOT and Kea Tijdens’s Wage Indicator Survey approach.

All presentations would be circulated to participants and submitted to the DASISH website.

1. See <http://www2.warwick.ac.uk/fac/soc/ier/software/cascot/> for further details and exemplar. [↑](#footnote-ref-1)