Using Wmatrix to classify open response survey data in the social sciences: observations and recommendations

Gill Philip
University of Macerata
gill.philip@unimc.it

Lorna J. Philip
University of Aberdeen
l.philip@abdn.ac.uk

Alistair E. Philip
Chartered clinical psychologist
aephilip@waitrose.com

1 Introduction

We report here on our use of Wmatrix (Rayson 2009) and the USAS tagger (Rayson et al. 2004) as an alternative to more commonly used content analysis methods for sorting and coding open response survey data in the social sciences.

Survey-based research in the social sciences often elicits open response data which is transcribed then sorted and coded, a procedure known as content analysis (Philip and Macmillan 2005). This methodological approach may be conducted by an individual researcher or by several members of a research team who then discuss their classifications to arrive at a final, definitive coding. Two particular problems arise. Firstly it is a time-consuming method, particularly in the preferred approach when more than one researcher participates in the exercise. Secondly, it is difficult to ascertain the accuracy and consistency of the coding *within* and *between* projects, i.e. in situations where more than one set of open responses in a single questionnaire need to be coded, and where similar topics are the focus of questionnaire based data collection in a number of projects. Replication is difficult because the number of categories and the level of detail that emerge from content analysis can vary considerably from one coder to the next and from one set of responses to the next. Having a finite, fixed set of categories would therefore be helpful, as would any degree of automation of the coding procedure. It is within this context that our experimentation with Wmatrix begins.

2 Extending Wmatrix to non-linear text

The decision to try out Wmatrix in the context of coding survey data was knowingly experimental. The program is designed to give its most reliable output in running text since determining the semantic class of a word is most effectively done when it is contextualised both semantically and syntactically (Rayson et al. 2004). We were well aware that the Wmatrix output might not be useful at all, because the type of data we were interested in – open responses to survey questions – comprises discrete words and short segments of text, but we thought it worth experimenting with in any case, since any tool which can considerably reduce the number of hours spent manually coding is potentially invaluable to social sciences research. Our default option, if the Wmatrix output were to prove unsatisfactory, was to use the USAS tagger to provide us with possible codes for our data, and to manually select the most appropriate one in the context. In the event, this was only necessary in three cases – to correct wrongly-coded words, to code wrongly-spelled words, and to supply codes for uncoded words (see Section 4). This was fortunate, because deciding which of several possible codes is the best fit is a time-consuming, sometimes frustrating business – possibly more time-consuming than assigning codes from scratch (see Section 5).

3 Word frequency and conceptual centrality

Our data are responses within a word association task. They appear to be fragmentary, but the words and short phrases for each section cohere at a cognitive level (this is true in general of open-response survey data).

Word association tasks are widely used in psychology and in some areas of linguistics, and request that participants state the first thing that comes to mind when they encounter a given probe word. Typically, those words (concepts) that are most centrally related to the probe word are mentioned first, with less central words/concepts appearing lower down the list, if at all. What the researcher hopes to find in the data is that all or most respondents will supply central words/concepts, while less central words/concepts will occur with much lower frequency and with greater lexical variety. What this means in practical terms is that a semantic core should make itself strongly visible due to the constant reiteration of central words/concepts, while the full extent of the semantic dispersal of the concept – which fields it touches on, and in what proportions – is informed by the less central words/concepts. There are evident parallels here with word frequency and collocation, except that in word association the co-occurrence phenomenon of interest is more abstract, something akin to Sinclair's (1996) semantic preference. The conceptual areas can be identified on the basis of raw frequency, after the semantic tags have been applied, but it is also interesting to apply a further test, since Wmatrix makes it possible for us to do so:

a comparison of the semantic fields in our data against the semantic fields found in the BNC for the same probe word (corpus search term). This allows us to highlight the semantic areas that are significantly present in our respondents' data compared to the language in general and is of particular interest to our ongoing main study because we want to assess students' vocabulary and conceptualisations of discipline-specific key words before and after taking a degree level course in Rural Geography – an area of study where lay and professional knowledge overlap and compete. [1] Comparing open response survey data with the normative data provided by the BNC is something that – to our knowledge – no previous studies of this type have attempted. This adds a further level of robustness to our qualitative analysis of data.

4 Manual intervention

The Z category in the USAS tagset is populated with grammatical words, proper names and unrecognized words (Rayson et al 2004). This is useful since it stops them from interfering in 'proper' text analysis using semantic tagging, where the focus is on semantic areas rather than structure. Our use of Wmatrix, however, is a little different from text analysis proper, since we are working with discrete words and short text fragments. It was therefore useful for our research to re-code the Z category tags wherever possible. In particular, we needed to look closely at:

- proper names with metonymical reference (e.g. 'Range Rover' standing for off-road vehicles and the people who drive them);
- proper names with restricted (local) meaning (e.g. 'King Street', specifically King Street in Aberdeen);
- acronyms (e.g. SEPA – Scottish Environmental Protection Agency);
- dialect and regional expressions (e.g. 'doofer', synonymous with thingamajig);
- archaic or non-standard spellings (e.g. 'fayre').

After dealing with these, we were left with what we are for now calling the 'Post-Office problem'.

5 The 'Post Office' problem

Wmatrix recognizes many compound nouns and codes them as single lexical items; but it does not know all compound nouns. *Post Office* – a recurring term in our data – was one of these. It had to be manually coded from the USAS tags for *post* and *office* respectively, resulting in a final coding as Q1.2 (paper documents and writing). Ideally, the code would have been for "services", but no such code is present in the tagset. We resisted the temptation to create a new class but remain not fully convinced of the choice made since it seems overly restrictive: as well as dealing with the delivery and reception of letters and parcels, post offices are retail outlets, offer a range of financial products and provide access to official services. In rural areas the post office van, until recently, was a mode of transport which allowed people to travel between places which were not served by public transport.

At the opposite end of the scale are words which attract a mind-boggling number of codes, none of which really seem to fit. USAS finds a total of 23 possible codes for *Costa* (in our data, the coffee shop), none of which captured the 'having coffee as a social event' sense expressed by the response 'Costa with friends' [probe: SOCIAL]. Such problematic items require discussion and debate before a definitive code is agreed upon.

6 Concluding remarks

We find that Wmatrix is an extremely useful tool for the initial coding of data such as that generated by open response survey questions, due largely to its speed of processing and its overall consistency and reliability. That said, we stress that it is essential to check all the output, not only to make sure that codes have been assigned correctly, but because compounds, phrases and, in some cases, even single-word responses, may benefit from multiple coding. Recurrent miscodings (not found in our data) or Z-category dumping (as in our 'Post Office problem') can often be resolved with reference to the USAS tagset. The USAS tagger is not fail-proof, however, and the researcher(s) conducting the analysis may need to make a fresh decision on the basis of the contextual cues of the response and the probe which it relates to. However a major benefit of Wmatrix is that it highlights semantic areas that could be easily overlooked because they are not central to the object of study, e.g. 'aesthetic judgement', and this further enhances the quality of the data analysis.

References

Philip, L.J. and Macmillan, D.C. 2005. "Exploring values, context and perceptions in contingent valuation studies: the CV Market Stall technique and willingness to pay for wildlife conservation". *Journal of Environmental Planning and Management* 48 (2): 257-274.

Philip, G., Philip L.J., and Philip, A.E. 2014. "Learning as conceptual acquisition: A pilot project for measuring learning outcomes in higher education." Paper presented at AELCO-SCOLA, Badajoz (Spain), 15-18

[1] Ethical approval for the study was obtained from the University of Aberdeen College of Physical Sciences Ethical Review Committee.

October 2014.

Pragglejaz group. 2007 “MIP: a Method for identifying metaphorically used words in discourse.” *Metaphor and Symbol* 22 (1): 1-39.

Rayson, P. (2009) Wmatrix corpus analysis and comparison tool. Computing Department, Lancaster University. http://ucrel.lancs.ac.uk/wmatrix

Rayson, P., Archer, D., Piao, S. L. and McEnery, T. 2004. “The UCREL semantic analysis system.” *LREC 2004*: 7-12.

Sinclair, J.M. 1996 “The search for units of meaning.” *Textus* 9: 75-106.

Steen, G., Dorst, A., Herrmann, J., Kaal, A., Krennmayr, T. and Pasma, T. 2010. *Finding Metaphor in Grammar and Usage.* Amsterdam: John Benjamins.