Linguistic-computing methods for analysing digital records of learning.

Richard Forsyth[1], Shaaron Ainsworth[1], David Clarke[1], Pat Brundell[1] and Claire O'Malley[1].

[1]**School of Psychology, University of Nottingham, NG7 2RD.**
Correspondence: rsf@psychology.nottingham.ac.uk

Abstract: Social scientists face an overload of digitized information. In particular, they must often spend inordinate amounts of time coding and analyzing transcribed speech. This paper describes a study, in the field of learning science, of the feasibility of semi-automatically coding and scoring verbal data. Transcripts from 48 individual learners comprising 2 separate data sets of 44,000 and 23,000 words were used as test domains for the investigation of three research questions: (1) how well can utterance-type codes assigned to text segments by humans be predicted from the linguistic characteristics of those text segments? (2) how well can learning outcomes be predicted from learners' verbalizations? (3) can the material students are learning from be identified from their language? Initial results indicate that the answers to the third question is yes; and that the answer to the first two questions is: well enough to warrant further development of the text-mining techniques so far employed.

Keywords: Applied linguistics, digital records, learning sciences, linguistic computing, machine learning, text mining.

1. Introduction

1.1 Background

Interaction is crucial to learning. People interact with fellow students, teachers, and -- increasingly -- various technologies such as computer-based learning environments and mobile devices. If these interactions are recorded, researchers can analyse them in order to understand the processes and outcomes of learning. Digital records are now routinely created of learners' experiences such as videos of individuals and groups, screen capture of actions on interfaces, system logs generated by the environments themselves, and audio recordings of talk by the learners. There is increasing interest in how these records might capture the breadth and depth of learners' experiences, to inform personal reflective learning, to support interventions from knowledgeable others, and even to replace formal assessment (e.g. CRA, 2005). However, the tools to automate the collection of such data have developed much faster than the tools to help us analyse the data. We can easily be swamped by potentially valuable data that we do not have the time or means to analyse fully. In this paper, we show how text-mining techniques could help learning scientists, and others, analyse one of the most important kinds of digital record -- transcribed speech.

To understand the processes involved in learning, we often need to analyse the content of learners' communication by coding each utterance according to the research question in hand (Ericsson & Simon, 1984). We may need to know the type of questions teachers are asking and how learners are responding (Dillon, 1982); whether learners are asking for help when they do not understand (e.g. Wood & Wood, 1999); or the breadth and depth of arguments they engage in (e.g. Andriessen, Baker & Suthers, 2004). When learners are alone, we need to know whether they speak or write explanations to themselves rather than simply paraphrasing material, as learners who do so learn more effectively (Chi, Bassok, Lewis, Reimann & Glaser, 1989).

Often this kind of coding is the single most time-consuming aspect of a study. Speech may have to be transcribed, and the text segmented. Coding schemes have to be formulated and then applied to each segment of text. Typically, at least two researchers code the data in order to assess reliability. Analysing an hour of verbal protocols can therefore take ten to fifty hours. So researchers often focus on just a few learners in the situation; or take a short sample; or work in more artificial situations where the interaction can be constrained. But we need to study processes over **longer** periods of time, and in considerable detail, if we are to understand how learning changes with experience (e.g. Siegler & Crowley, 1991). We also need to understand how learning occurs in real contexts, not just artificial ones (e.g. Cobb et al, 2003).

1.2 Research Questions

Our overall aim is to assess the extent to which modern text-mining techniques can help researchers deal with digital records produced in learning situations. In the present study this gave rise to three specific questions:

> (1) [Categorization] how accurately can an automatic system, using linguistic information, assign functional codes to transcribed speech segments?
>
> (2) [Outcome Prediction] how well can computational analyses of learners' language predict learning outcomes (measured by pre- and post-test scores)?
>
> (3) [Context Identification] can the material students are learning from be identified from their language?

If such tasks can be done automatically or semi-automatically, and accurately, we should have largely overcome a major obstacle to progress in this field, namely the laboriousness of encoding transcripts of learners' speech. This would also benefit researchers in other fields where large volumes of text need to be coded for analysis.

2. Technical Details

From over 20 widely available linguistic processing systems, five were judged promising for the task in hand. In this paper we concentrate chiefly on results obtained using the WMatrix package (Rayson, 2005). Our representation scheme is based on a simple document model: a corpus is a collection of texts; a text is a sequence of segments; a segment is a sequence of bytes/characters.

3. Initial Datasets

Our starting point was a self-explanation study by Ainsworth & Burcham (2004). Self-explanations are pieces of knowledge generated by an individual learner that state something which is not explicit in the information they are learning from (Chi et al, 1989). This is of interest because learners develop a deeper understanding of the material they are studying if they give more self-explanations.

The Ainsworth & Burcham (2004) study produced transcripts of 24 learners studying the cardio-vascular system for roughly one hour, reading text that was either high or low in coherence. A second self-explanation study was also analyzed in which learners studied either abstract or realistic diagrams (Robertson, 2004). The transcripts were segmented and coded according to the nature of the utterance (paraphrase, self-explanation statement, or monitoring statement). In addition, pre- and post-tests of learners' understanding of the cardio-vascular system were conducted, and shown to relate to self-explanation behaviour. It took an estimated twenty hours effort for each hour of student learning to produce these data. Details are given in Table I.

Table I -- Dataset Details.

Dataset:	AB (Ainsworth & Burcham, 2004)	AR (Robertson, 2004)
Participants:	24 (13 female, 11 male)	24 (13 female, 11 male)
Words:	44,388	23,330
Segments:	2071	1784
Material:	text	diagrams
Conditions:	high versus low coherence	abstract versus realistic
Assessment:	3 pre-test & 5 post-test measures	3 pre-test & 5 post-test measures

These data provide an ideal starting point for considering the issues raised in this paper: they are time-consuming for researchers to code, yet relatively simple linguistically (monologues in artificial situations analysed according to well a established coding scheme), and there is widespread interest in this type of coding.

4. Results

Table II -- WMatrix Tagging Output (AR27; 35).

```
    0000042 010  CC      And               Z5
    0000042 020  PPH1    it                Z8
    0000042 021  VBZ     's                Z5 A3+
    0000042 030  VVG     showing           A10+ S1.1.1
    0000042 040  RR      obviously         Z4 A11.2+
    0000042 050  AT      the               Z5
    0000042 060  JJ      main              A11.1+ N5+++
    0000042 070  NN1     pump              O2 B5
    0000042 080  VBZ     is                A3+ Z5
    0000042 090  AT      the               Z5
    0000042 100  NN1     heart             B1 M6 A11.1+ E1 X5.2+
```

Initially, WMatrix was applied to all 48 texts, to obtain semantic and syntactic tagging data. Specimen WMatrix output is shown as Table II. Syntactic codes are in column 3 and semantic codes in column 5 (starting with the program's "first-choice" code).

4.1 Segment Categorization

Text segments were coded into 3 higher-order categories; thus each text segment could belong to one of three classes: paraphrases, self-explanations and monitoring statements. For example, if a student read "The septum divides the heart lengthwise into two sides", a paraphrase might be "The septum is what goes down the middle of the heart"; a self-explanation might be "Septum is what separates the two ... some sort of control", and a monitoring statement might be "I'm not sure why".

A linear classifier was applied to this data 6 times -- 2 data sets using 3 different groups of variables: Counts (simple frequencies of each linguistic feature), Rates (counts divided by segment size), and Transitions (digram conditional probabilities).

Assessing the quality of a multi-class categorizer is not trivial. Raw success rate (or error rate) is misleading when the classes are skewed in frequency, as here. In this study we used a proportional reduction in error (PRIE) statistic, equivalent to the tau coefficient (Klecka, 1980), with the chance proportion of success calculated according to a formula derived from Mosteller & Bush (1954). This gives the reduction in the rate of errors compared to guesswork. It is analogous to the proportion of variance accounted for in a regression. The results obtained with a stepwise procedure using a maximum of 8 variables are tabulated in Table III. These are cross-validated scores, which are not optimistically biased.

Table III -- PRIE scores of Linear Classifier in 6 conditions.

Dataset:	AB	AR
Counts:	27.57%	39.46%
Rates:	28.71%	35.98%
Transitions:	27.19%	21.48%

The transition data performs worst in both data sets. Surprisingly, the shorter segments of the AR dataset (median 12 words) are more accurately classified than those of the AB dataset (median 19 words). This may be because the segmentation was less precise in the AB dataset, leaving a number of segments compounded of two different utterance types. The best result (AR dataset with Counts) was obtained using 6 syntactic variables and 2 semantic variables.

4.2 Learning Outcomes

First a Principal Components Analysis was applied to reduce the five post-test scores to 2 major dimensions accounting for 50%+24% of the total variance (AB data); or 55%+21% of the variance (AR data). The first dimension was clearly an overall "success" score. This was used as the DV in a number of regression studies, using multiple linear regression and regression trees (Venables & Ripley, 1997).

With only 24 cases and over 200 features, a severe overfitting problem presents itself. It is possible to achieve an apparent R-squared of more than 99% by using 20 or more predictors. Side studies with this data indicated that the optimum number of variables in a linear regression was 3 and that the best number of leaf nodes to allow in a regression tree was also 3. These limits are used in what follows.

The best 3-variable linear regression used high-frequency vocabulary items, not syntactic or semantic tags, generating the following formula (adjusted $R^2 = 0.69$):

$$Dim1 = 0.722 + 0.105*AVOCbody -64.761*RVOCit -0.040*AVOCof$$

Here AVOCbody and AVOCof are the absolute counts of the words "body" and "of", while RVOCit is the relative frequency of "it". Note that "it" and "of" have negative coefficients, meaning that lower frequencies predict more successful learning, whereas higher frequencies of the word "body" predict more successful learning.

The best 3-leaf-node regression tree was grown on the AB data. It is shown in Figure 1. This used semantic variables from WMatrix, Rsem.Z8 and Asem.Z9. The former is a relative frequency and the latter an absolute frequency. This raises an issue of "back translation". To help interpret the tree it should be noted that values larger than the given threshold always take the right branch (and smaller the left). Table IV shows the 8 most common words from the AB corpus belonging to these two categories.

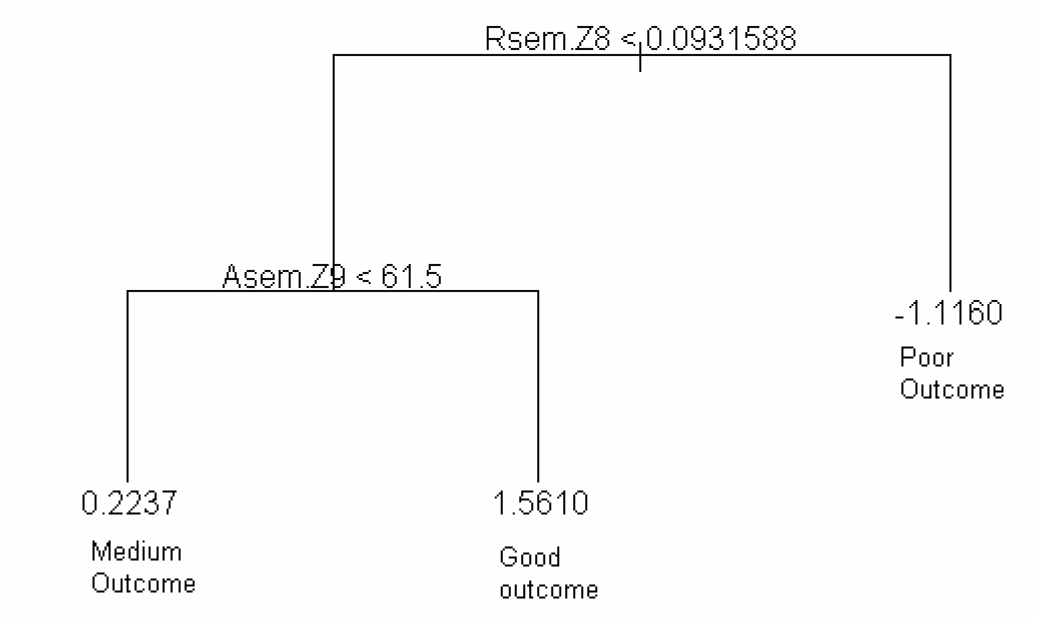

Figure 1 -- Regression tree using semantic tags (AB dataset).

Table IV -- Most frequent words in AB corpus associated with tags Z8 and Z9.

WMatrix semantic category	**Most common words in AB corpus receiving that tag**
Rsem.Z8 [pronouns: relative frequency]	it, that, they, you, which, i, this, we
Asem.Z9 [technical terms: absolute freq]	deoxygenated, oxygenated, arterioles, tricuspid, carbon-dioxide, venules, systole, bicuspid

4.4 Source-Material Identification

A stepwise linear discriminant function using up to four linguistic variables was able to classify the AB texts as coming from the High or Low coherence condition with 96% cross-validated success rate (100% using resubstitution), and to classify the AR texts as from the Abstract or Realistic condition with the same success rates. Clearly the source material does leave a mark on the language used by participants. Indeed, with the AR data, it is possible to achieve near-complete linear separation with only 2 variables (occurrence rates of the words "pumped" and "ventricles"), as in Figure 2.

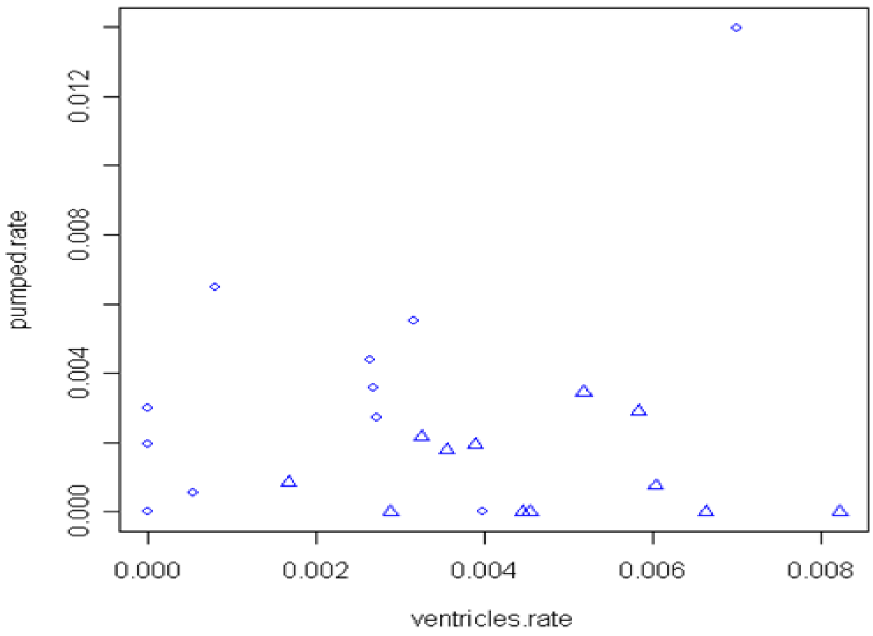

Figure 2 -- Separating conditions, AR data (blobs=Abstract, triangles=Realistic).

5. Discussion

These early results show considerable promise. In two different data sets the imprint of the experimental conditions could be clearly detected in the language used by participants. More importantly, short text segments averaging only a dozen or so words could be categorized using a combination of linguistic features. Moreover, it is possible to account for a large proportion of the variance in learning-outcome scores using predictive formulae based on linguistic features. Such findings hold the promise of further insight into the process whereby self-explanation affects learning success.

If these early results prove robust, this approach could provide a foundation for convenient analysis of naturally-occurring data from collaborative learning situations which have previously been too time-consuming to code, and open up similar opportunities for researchers in numerous other fields where text-coding is a problem. We believe it is worth pursuing this approach further, employing a wider range of linguistic markers as well as more sophisticated algorithms.

References

Ainsworth, S.E & Burcham, S. (2004) Limits on the Self Explanation Effect. Paper presented at the *EARLI SIG2 Meeting* Valencia, September 2004

Andriessen, J., Baker, M., & Suthers, D. (Eds.). (2004). *Arguing to Learn: Confronting Cognitions in Computer-Supported Collaborative Learning Environments*. Amsterdam: Kluwer Academic Publishers.

Chi, M. T. H., Bassok, M., Lewis, M. W., Reimann, P., & Glaser, R. (1989). Self-explanations: How students study and use examples in learning to solve problems. *Cognitive Science*, 5, 145-182.

CRA (2005). Cyberinfrastructure for Education and Learning for the Future: a Vision and Research Agenda 2005. Computing Research Association.

Cobb, P., Confre, J., diSessa, A., Lehrer, R., & Schauble, L. (2003). Design experiments in education. *Educational Researcher, 32*(1), 9-13.

Dillon, J. T. (1982). The Multidisciplinary Study of Questioning. *Journal of Educational Psychology*, 74(2), 147-165.

Ericsson, K. A., & Simon, H. A. (1984). *Protocol analysis: verbal reports as data*. Cambridge, MA: MIT Press.

Klecka, W.R. (1980). *Discriminant analysis*. Sage Publications, Newbury Park, California.

Mosteller, F. & Bush, R.R. (1954). Selected quantitative techniques. In G. Lindzey (ed.) *The Handbook of Social Psychology*, Vol.1, Addison-Wesley, Reading, Mass. 289-334.

Rayson, P. (2005). *Wmatrix: a web-based corpus processing environment*, Computing Department, Lancaster University. http://www.comp.lancs.ac.uk/ucrel/wmatrix

Robertson, L.A. (2004). *Learning with diagrams: the effects on learning by self-explanation with perceptually realistic and abstract diagrams*. Unpublished BSc dissertation, University of Nottingham.

Siegler, R. S., & Crowley, K. (1991). The microgenetic method - a direct means for studying cognitive-development. *American Psychologist*, 46(6), 606-620.

Venables, W.N. & Ripley, B.D. (1997). *Modern applied statistics with S-Plus*, 2nd edition. Springer-Verlag, New York.

Wood, D. J., & Wood, H. A. (1999). Help seeking, learning and contingent tutoring. *Computers and Education*, 33(2/3), 153-1770.