WHAT STUDENTS REALLY SAY

Mary McGee Wood, John Sargeant & Craig Jones

What Students Really Say

Mary McGee Wood, John Sargeant, & Craig Jones
mary, johns, cjones@cs.man.ac.uk

Introduction

The Assess By Computer (ABC) project (Sargeant et al 2004) follows a Human-Computer Collaborative (HCC) approach to assessment. We focus on constructed answers such as text and diagrams rather than answers requiring mere selection between alternatives. The HCC assessment process is an active collaboration between humans and a software system, where the software does the routine work and the humans make the important judgements. Similar approaches in Artificial Intelligence research are developed in Englebart 1962, Grosz 2004, and Potter et al 2004, among others.

Our students, through their answers to questions, also implicitly collaborate in the development of resources. We can develop marking support tools which handle the nature and range of variation found in real exam data, and we can adapt marking judgements and feedback - even, in the longer term, our teaching material - in the light of what students *really* say.

In this paper we focus on the reality of student text answers. We present student data from on-line examinations showing a remarkably wide range of acceptable answers to even the most straightforward of questions. We show how the analysis of these examples is supported by the ABC tools, especially the Keyword Manager and answer clustering options.

What students really type

Real student answers are often acceptable, while not strictly "correct". Mis-spelllings are all too often found in genuine exam data (as we have discussed previously, Sargeant et al 2004), and there is no chance of "benefit of the doubt" in typed answers: typescript is remorselessly legible. An extreme example is shown in figure 1.

Two other important types of acceptable answers beyond the pre-specified "model answer" are word variants, and context-dependent synonyms. The latter was also discussed in our previous paper; the former is addressed here.

Word variants

Word variants are minor, context-independent alternatives - notations or forms of a word which differ from the "model answer", but are still "correct". Their

range turns out to be surprisingly wide even for straightforward short factual answers, and variable in their amenity to reliable automatic detection.

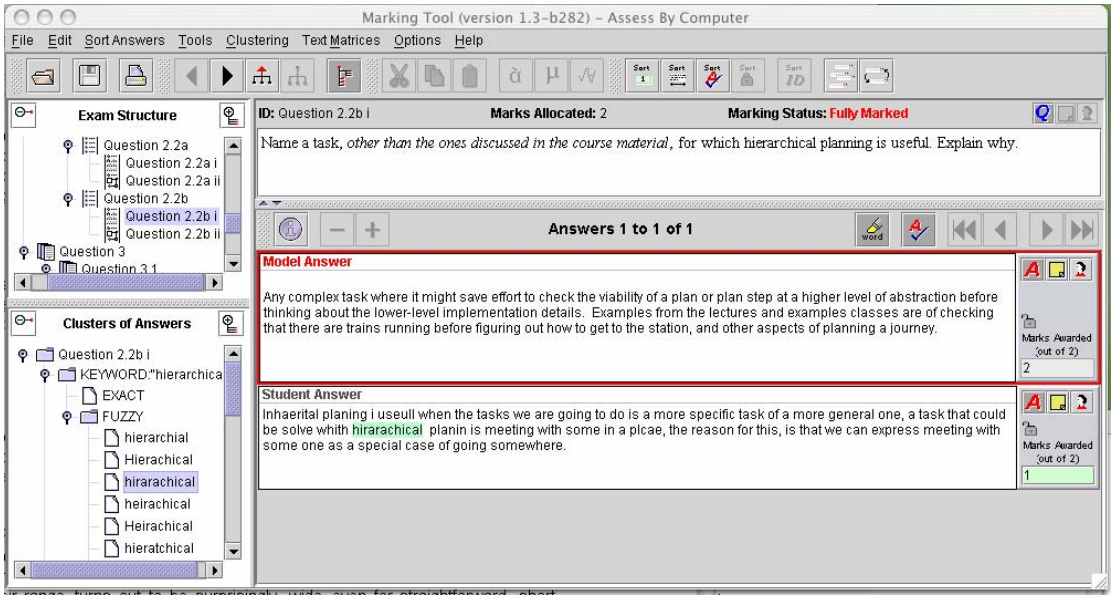

Figure 1. An unusual mis-spelling of the word “hierarchical”

In what ought to be a trivial example from a first-year undergraduate exam in Artificial Intelligence, a problem on probability was set to which the "model answer" was "P (A1) =1/2". 114 students answered the question, and produced 21 distinct answers, 13 of them correct and 8 incorrect. There were 49 instances of "1/2" (43%), but a very long tail, with 13 singletons, 5 of them wrong. The variants are easily dealt with (e.g. 1/2, 0.5, .5, 0.50, 50%), but it is still instructive to see how many different ways a group of students can find to say something very simple.

A second, more challenging example comes from an open-book, untimed test in human biology including the question "What is haemolytic disease of the newborn? How can this be prevented?" A critical answer key phrase was "rhesus positive". The 281 student responses produced 52 distinct ways of expressing this phrase. 15 were mis-spellings, leaving 37 distinct correct representations.

Analysis revealed six parameters of variation:

Upper / lower case: rhesus / Rhesus / RHESUS

Hyphenation: RH-positive / Rh positive

Spacing: Rh +ve / RH+ve

Parentheses: +ve / (+ve)

"D": Rh positive / Rh D positive

"positive": positive / pos. / pos / + / +ve / +ive

Clearly this defines a far larger space of possible acceptable variants (288) than the 37 found already (it will be interesting to see how many new ones turn up in next year's repeat of the test). Mis-spelings must be dealt with as well, and allowing an edit distance of even one will allow "RH-" - precisely the *wrong* answer.

If this nature and degree of variation is found even for an objective two-word phrase in a technical domain, in an open-book test with no time pressure, it is somewhat alarming to speculate what we may find when we begin to look at less constrained situations, such as (for example) language translation exercises.

What students really draw

The ABC exam client also allows students to draw simple diagrams. Work on graph matching (Tselonis et al 2005) is addressing the usual, objective case where we want student answers to match a pre-specified answer, such as the correct representation of a molecule in chemistry. Diagrams in subjective, open-ended answers pose a different set of problems.

Partial order planning is an AI technique which separates independent sequences of processes within a complex task. Students were asked to draw a diagram representing an original example of this. The first example (figure 2) is exactly right. The second (figure 3), although it has the right shape, is exactly wrong: the whole point of the technique is to avoid enumerating all the possible orders of actions. The third (figure 4), although graphically completely different, is conceptually correct and received full marks.

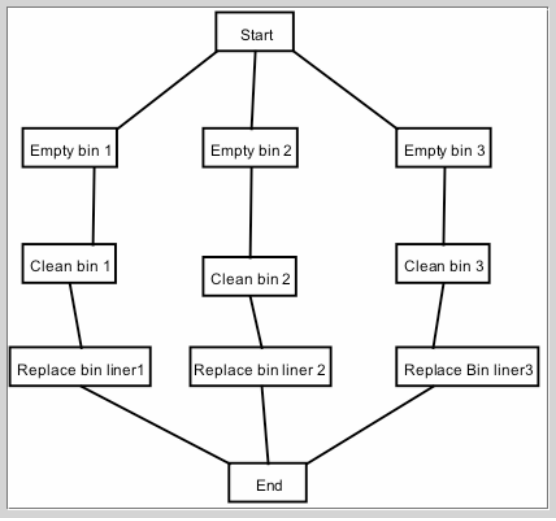

Figure 2. Correct example of a diagram representing partial order planning

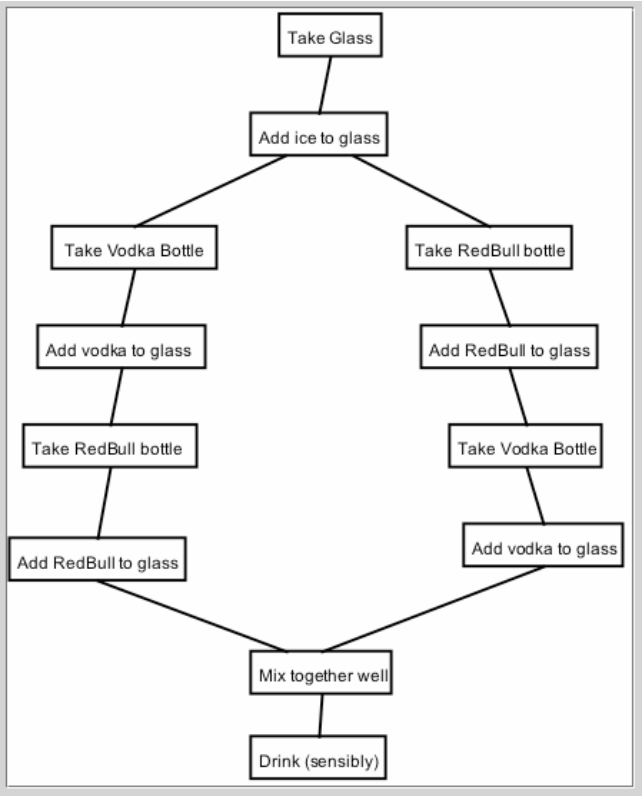

Figure 3. Incorrect example of a diagram representing partial order planning

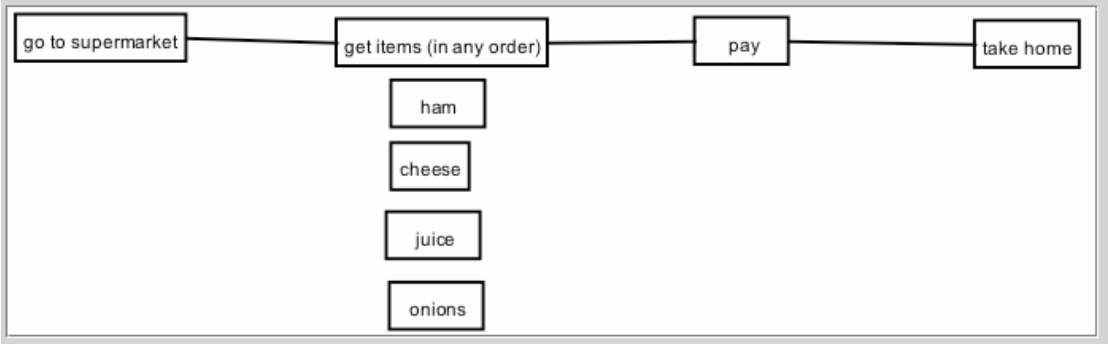

Figure 4. An alternate correct example of a diagram representing partial order planning

What students really know

What are students telling us, through their input in assessments, about what they have learned?

Context-dependent resolution of ambiguity

104 students answered the question "In Artificial Intelligence, what is the "frame problem"? The model answer was "The search spaces for real world problems are too large to re-compute completely when something changes. `Only compute what has changed in the situation.' " However the real point of the question was to find out how many of the students had understood the fact - stressed in the lectures - that the "frame problem" has nothing to do with "frames" as a knowledge representation formalism, discussed in another part

of the course. In other words, could they correctly resolve the ambiguity in the word "frame", given the context "problem"?

Of 40 answers which received the maximum two marks, all but two contained at least one of the keyword family "change / changed / changing": both the exceptions were untypically long, and included original ideas. Of 38 answers which received no marks, all but three contained none of these keywords: those three did contain keywords such as "inheritance" which reliably identify the wrong meaning of "frame".

Although it was predicted that words like "slot", "value", and "inheritance" would be useful indicators of wrong answers, in practice the students who were weak enough to make the mistake tended also to lack the correct vocabulary to make it with: only five answers used the word "inheritance", for example. The unpredicted useful keyword was "frames", plural. All 13 answers which contained this and none of the "change" set were wrong. Another eight containing both were given one or two marks each.

Original answers

In the previous cases, there was a correct answer, although it could be expressed in a surprising number of different ways: we call these “objective” questions. “Subjective” questions - open-ended questions with no one, pre-set correct answer - are a different story.[1]

The example given here comes again from a first-year exam in Artificial Intelligence. The question was "Give an original example of an exception to default inheritance"; the example used in the lectures was of penguins, which are birds but do not fly. 109 student answers provided 122 acceptable distinct examples. The vast majority are singletons: the interesting - and disappointing - thing about them is how similar most of the "original" answers are to the exemplar.

Despite the explicit requirement for an original example, nine students cited penguins as non-flying birds. This was followed in frequency by six ostriches, and a total of 20 non-flying birds, over 16% of all example tokens given. There were eight non-walking mammals (four bats, three whales, and one dolphin), a further 30 naturally anomalous animals, and 31 disabled or injured animals (including blind and bald humans). In total some 73% of all answer tokens - 89 - involved animals, as against five plants and 28 artefacts.

Looking more closely at the artefacts, we find that the more original answers - the answers least similar to the model - are often among the best, while a few are among the worst. The latter include cars without radios and trousers without pockets: bad examples because one would not think of radios and

[1] Our terminology here differs from that of the CAA community at large, who use “objective" to mean selected. We consider that the objective / subjective and constructed / selected distinctions are largely independent of each other. We further believe that the distinction between “long” and “short” text answers (e.g. Leacock & Chodorow 2003) is relatively superficial.

pockets as central, prototypical, defining characteristics of cars and trousers in the way that flight is for birds. The former include:

- Woodwind instruments have reeds, except for flutes, which do not.
- Aircraft have wingspan measurements, except for helicopters, which do not.

This is to be expected, as a truly original answer is a better indicator of understanding (or the lack of it) than one which is closely modelled on the exemplar.

From a pedagogic point of view, the lack of real originality is perhaps disappointing. (A conceptually identical question was set the following year using the example of white cricket balls: the answers this time included four orange or yellow footballs and a table tennis ball.)

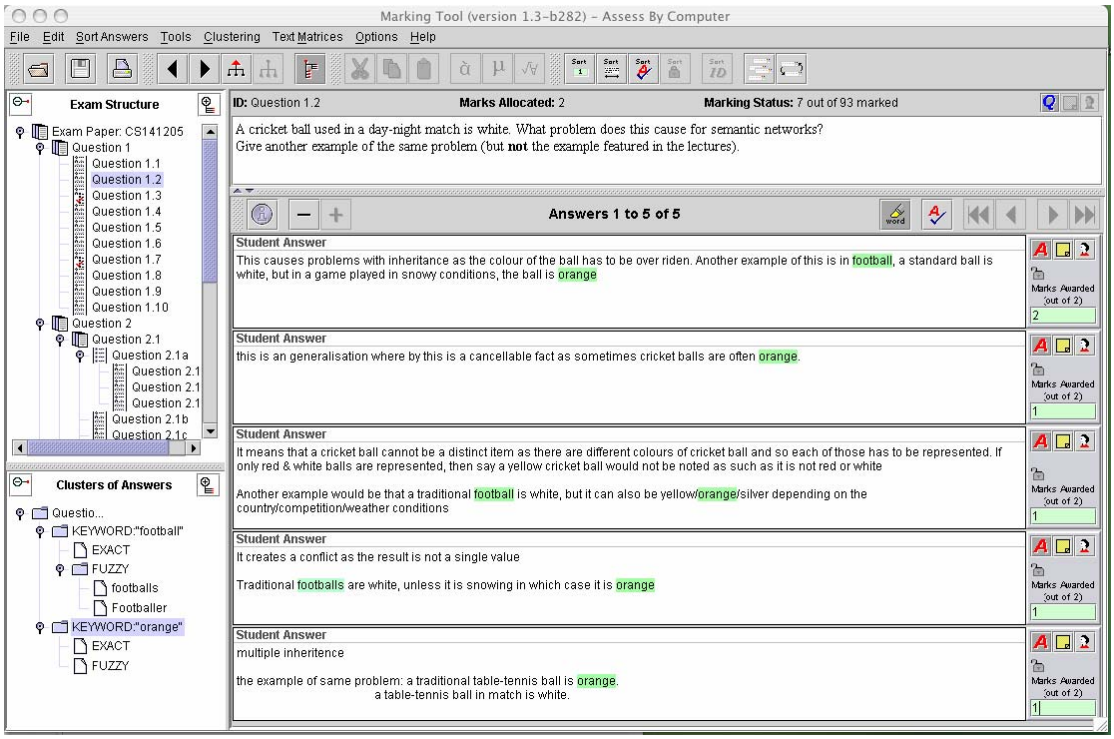

Figure 5. Answers displaying a lack of originality

The failure to discriminate between exceptions which are systematic properties of species or classes, and those which arise through accidents to individuals, was an unexpected finding. With hindsight, it is understandable, and not the students' fault: the first author, who set the exam, has amended her lecture material since.

What can we do about it?

One obvious way of making this large messy space of possibilities amenable to automatic analysis is to tightly constrain the form of student input. However, although this may be appropriate in some situations, it is pedagogically

unsuitable in others. For example, in the case of calculations - even if straightforward - one might well want to see the working as well as the answer. The answer could be tightly constrained to be a number typed into a small box; the working can be loosely constrained, but must leave space for individual variation. The form of student answers should be dictated by pedagogic motivation, not by the limits of the software.

The examples given support our philosophy that, in an HCC framework, simple tools can be effective in supporting a human marker. The range of variation is wide and difficult to predict; therefore, tools which support dynamic revision of Marking Judgement Representations on-the-fly are important.

The ABC marking tool offers a range of options for sorting answers: by anonymous number, answer length, marking status, or similarity to model answer based on keywords. Its dynamic Keyword Manager specifies keywords to be highlighted in answers, with optional fuzzy-matching. Keywords can be added or removed on-the-fly, so any unexpected frequent words discovered during the marking process can be added. The analysis of original answers given above was achieved very quickly by moving a succession of animals in and out of the Keyword Manager and re-grouping by similarity. The analysis of the “frame” ambiguity used an interleaving of fuzzy-matching and keyword grouping, and again took very little time.

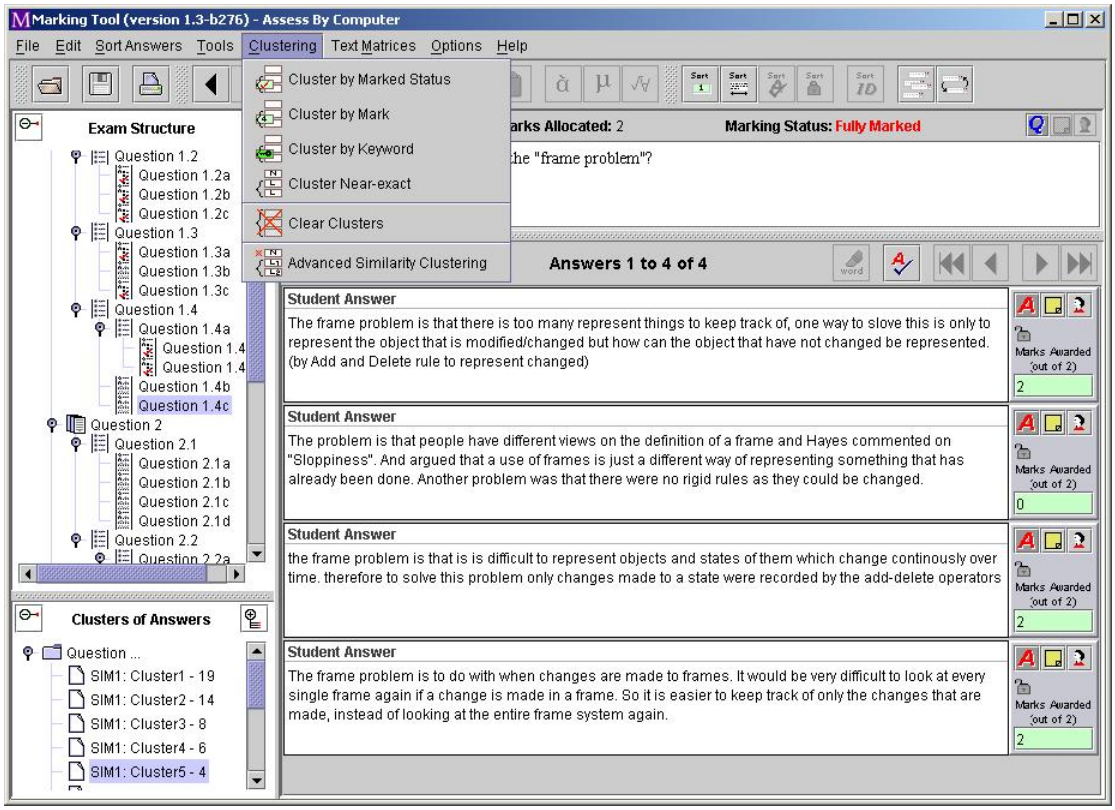

Figure 6. Enhanced clustering techniques in operation

In addition to grouping by specified keywords, work by the third author has enhanced the ABC marking tool (figure 6) with robust, generic text clustering techniques drawn from Natural Language Engineering (Jain et al 1999).

These techniques, used elsewhere in automatic summative marking (Shermis & Burstein 2003), are used here to group together similar answers to improve the speed and accuracy of the human marking process.

Why does it matter?

The ABC tools are being developed incrementally, informed by the patterns we find in the reality of student text answers. There seems to be an (understandable) tendency in CAA and other analogous domains (e.g. the development of automatic Tutorial Dialogue Systems: cf Wood 2005) to base hopes and research on unrealistic assumptions about the nature of real data. We believe we have demonstrated that techniques developed without proper consideration of real data are inherently flawed. Our aim is to develop realistic assessment support tools, based on real data, which help the human marker efficiently and consistently to evaluate what students *really* say.

References

Engelbart, D.C. Augmenting Human Intellect: A Conceptual Framework. Summary Report, Stanford Research Institute, 1962.

Grosz, B.J. Beyond Mice and Menus. Proceedings of the American Philosophical Society 2004.

Jain, A.K., Murty, M.N., Flynn, P.J. Data Clustering: A Review. ACM Computing Surveys 31:3 1999.

Leacock, C., Chodorow, M. C-rater: Automated Scoring of Short-Answer Questions. Computers and the Humanities 35:4 2003.

Potter, S., Tate, A., Dalton, J. Collaborative task support and e-Response. AISB Quarterly 115 2004.

Sargeant J., Wood M.M., Anderson S.M.: A human-computer collaborative approach to the marking of free text answers. Eighth International CAA Conference, Loughborough University, Loughborough, UK, 2004 pp.361-370.

Shermis & Burstein (eds) Automated Essay Scoring, a cross-disciplinary approach, Lawrence Erlbaum 2003.

Tselonis, C., Sargeant, J., Wood, M.M.: xxx Ninth International CAA Conference, Loughborough University, Loughborough, UK, 2005.

Wood, M.M.: The Uses of Code Fragments in Programming Language Tutorials. Workshop on Mixed Language Explanations in Learning Environments, AIED, Amsterdam 2005.

Acknowledgements

We are grateful to the University of Manchester Distributed Learning Fund for their financial support which has made this work possible; to Stuart Anderson for building the original software; and to Katherine Getao for her insightful, challenging questions on earlier drafts.