ISSN 1744-1986

Technical Report N° 2007/ 18

Evaluation and Improvement of an Automatic Marking Tool for Diagrams

Martin Huw Ball

29 September, 2007

Department of Computing
Faculty of Mathematics, Computing and Technology
The Open University

Walton Hall, Milton Keynes, MK7 6AA
United Kingdom

http://computing.open.ac.uk

Evaluation and Improvement of an Automatic Marking Tool for Diagrams

A dissertation submitted in partial fulfilment
of the requirements for the Open University's
Master of Science Degree
in Computing for Commerce and Industry

Martin Huw Ball
(M9437754)

3 March 2008

Word Count: 16308

Preface

I wish to thank the many people and organisations who have helped me produce this dissertation. Thanks go to:

- My employer, IBM (United Kingdom) Ltd for financially supporting this project.
- Dr Andrea Sherriff for help and advice on statistical methods.
- Special thanks to my Open University supervisor, Lindsay Hughes for his helpful advice, encouragement and support throughout the entire project duration.
- The Open University research team of Dr Neil Smith, Dr Pete Thomas and Dr Kevin Waugh who produced the marking tool software and documentation. Special thanks to Dr Pete Thomas who, as Specialist Adviser for this project, provided help and feedback throughout the project.
- Friends and family for acting as proofreaders and for providing encouragement and support.

Table of Contents

List of Figures

List of Tables

Abstract

Within the teaching environment, the ability to automatically grade student assignments and examination answers offers many potential advantages, including improving the speed and standardisation of the marking process, and the release of time spent marking by tutors which can then be utilised for teaching.

A marking tool has been developed by the Open University as an investigation into the feasibility of automating the grading of student Entity Relationship (E-R) diagram answers. Two small scale trials of the marking tool were performed by the developers. These showed that automated marking was potentially feasible, and that detailed failure analysis of the incorrectly marked diagrams could be used to develop and implement improvements to the marking tool.

The research described in this dissertation extends the work of this team, using 2 new corpora of graded E-R diagrams. The first of these contains 197 diagrams and the second contains 32 diagrams. The second corpus introduced additional complexity as the model E-R diagram answer used entity subtypes within the solution diagram.

This dissertation describes the work undertaken to establish an experimental procedure to grade the diagrams using the marking tool; the establishment of statistical and graphical techniques to measure the performance of the marking tool; the root cause analysis of marking tool deficiencies where diagrams are incorrectly graded; and the design, implementation and testing of improvements to the marking tool.

Chapter 1 Introduction

Since the advent of the information age, significant research effort has been made in the use of technology to improve the teaching and learning processes. This work has been categorised under Educational Technology. A major topic within this subject is Computer Aided Assessment (CAA) (Bull, 2002; Warburton and Conole, 2003).

Automated marking systems have received significant attention as they have potential to reduce teacher workload and to improve the quality of education. However, very little research has been published on the methods that should be employed to evaluate the performance of automated marking systems, or the strategies or methods to improve the marking system performance.

1.1 Question Categories and Marking Tool Availability

Both Carter et al. (2003) and Tsintsifas (2002) document the availability of automated marking tools for different question types. Tsintsifas categorizes these into fixed and free response methods as below:

Fixed Response

- Multiple-choice questions.
- Simple text or numeric answer questions.
- Hotspot graphical questions.

Free Response

- Essay exercises.

- Computer programming exercises.

- Diagrams and graphics.

By their nature, fixed response marking tools are relatively simple to implement and give an accurate mark. These factors make them the most widely used automated tools.

In contrast, tools for free response marking are complex, and those documented are generally prototypes rather than fully validated offerings. A number of essay marking systems have been described. E-rater® (Attali and Burstein, 2005) is typical of these in that its success in consistently matching a human mark is limited.

Tools for marking computer programming exercises are the exception in that a number of successful implementations have been reported. The Ceilidh system described by Tsintsifas (2002) is an example.

For the automated marking of diagrams and graphics, several tools have been documented. The tool described by Thomas et al. (2006b), which is the subject of this thesis, is the most promising that I found.

1.2 Benefits and Disadvantages of Automated Marking

Carter et al. (2003) and Tsintsifas (2002) both discuss the benefits achievable using automated marking. A summary of these follows:

1. Time spent marking student assignments is reduced and leads to more efficient teaching allowing larger class sizes. Saved time can be used for teaching.

2. Automated assessment is well suited to distance learning.

3. Assessments can be made more frequently, assessing smaller chunks of the coursework.

4. Marking is performed quickly and consistently.

5. Consistent high quality feedback can be provided.

6. CAA can help with both the detection and prevention of plagiarism. Detection can be implemented with the automated comparison of student answers. Plagiarism can be deterred by presenting each student with randomly selected problem sets.

Carter et al. (2003) presented case studies on CAA showing the following perceived disadvantages:

1. Significant resources are needed in the setup stage of a new system and for question generation.

2. The unforgiving marking of which a computer is capable.

3. The restricted style of questions that can be electronically marked.

The first disadvantage is self-explanatory.

For the second disadvantage, Carter et al. explained that some tools "… *cannot distinguish between 'nearly correct' and 'wrong'*". An illustration of this disadvantage is the situation where a question has to be solved using a number of steps. Common marking practice is to deduct only a small penalty for an error in the early stages if the subsequent reasoning is correct. It is difficult to implement this marking practice in an automated marking tool.

The last disadvantage is a reflection of the lack of tools to support marking of free response questions, and the deeper concern that there is a lack of sound theoretical basis on which to construct these tools.

The following section explains why free response questions are important to educators.

1.3 Testing of Higher Cognitive Aspects of Learning

Tsintsifas (2002) described Bloom's taxonomy which divides the learning process into cognitive aspects that are ordered. At the lower level is knowledge. To assess, the assessment will test the student's ability to remember facts. The complete list of cognitive aspects within Bloom's taxonomy is shown in Table 1.1.

Cognitive Aspect of Learning	Ability
Knowledge	To remember
Comprehension	To understand
Application	To apply concepts to solve problems
Analysis	To break down to concepts
Synthesis	To combine concepts
Evaluation	To make judgements

Table 1.1: Cognitive aspects of learning according to Bloom (1956)

Tsintsifas asserted that all cognitive levels of Bloom's taxonomy are not easy to assess, if at all, using automatic assessment. For disciplines requiring design skills, it is much harder to devise automated assessment based on fixed response. Carter et al. (2005) provided evidence that multiple choice questions have been used to assess all cognitive levels except for synthesis. However, they acknowledged that multiple choice questions

are most suited to testing the lower cognitive levels. Carter et al. also introduced their own level, problem solving, which they said requires both analysis and synthesis skills.

My conclusions are that:

- Fixed response questions are most suited for lower level (pre-university and undergraduate) education levels where the lower cognitive level skills are to be tested. For these cases, automated marking tools are available and are realistic to implement.

- Free response questions are needed to test problem solving, synthesis and design skills. It is also easier to devise free response questions to test analysis and evaluation skills, as opposed to using fixed response questions. With the exception of computer programming exercises, there is little support available for the automated marking of these questions.

- The potential benefits of automated marking justify research effort into tools for the marking of free response questions.

1.4 Background to the Automated Marking Tool for ER-Diagrams

Thomas (2004a); Thomas (2004b); Smith et al. (2004) and Waugh et al. (2004), together provide a background and introduction to the automated marking tool for diagrams. They described the problem that within student diagrammatic answers, imprecise and incomplete diagrams convey some meaning and need interpretation. A five-stage architecture was described to resolve this issue. Diagram answers are first decomposed into a set of primitives. These are then identified as diagrammatic features and then aggregated into higher-level features and interpreted.

A series of experiments were performed on student diagram answers and the grades produced by the tool are compared to human grades. Mean and standard deviation are compared for each set of results and correlation data is produced using Pearson, Spearman and Kendall correlation methods (Kirkwood and Sterne, 2003). A problem with tool marking of low scores was identified and corrected, and tests were repeated. Overall correlation results showed promise.

Thomas et al. (2005) continued with further experiments using the tool on a question that was more open-ended. This exposed a further limitation of the tool when students had used synonyms for entity names. An additional stage was added to the marking algorithm to correct for this.

Additional details on diagram interpretation using patterns and sub-diagrams were described by Thomas et al. (2006a). The principles of the aggregation stage were described. Also, an explanation was given of how the problem of exact and inexact matching was overcome by using a similarity matrix. An example of this was given.

In their most recent publication, Thomas et al. (2006b) described their latest experimentation using the marking tool. Two new issues were investigated:

- For some questions, poor diagram layout had adversely influenced the human mark. Suspected incorrect human marking was reassessed to correct.

- For some diagram answers obtaining a lower mark, there was a larger discrepancy between the human and tool mark. It was found that the tool was finding matches between the model answer and the student's answer that a human would not consider to be reasonable. Detail was given on 2 new rules introduced to the tool to correct for this issue.

Thomas et al. (2006b) concluded that there was a need to extend their tools and techniques to a larger set of student diagrams.

1.5 Aims and objectives of the research project

The work of Thomas et al. (2006b) has demonstrated using 2 small scale trials that automated grading of Entity-Relationship (E-R) diagrams is potentially feasible using the tool that they have developed. The aim of this research project was to extend this work.

Firstly, I aimed to extend the second trial described by Thomas et al. by testing more student answer samples. Within their second trial, 14 student answers were marked using the tool. At the start of this project there were 2 corpora of data available, each relating to a question that has been used by The Open University within their undergraduate database course. Each corpus has one or more standard answers which the marking tool uses in order to compare the student answer and assign a score. The two corpora together contain approximately 260 student diagrams with an associated human mark which were used to test the tool's marking performance.

Secondly, I aimed to use this new information to make reasoned modifications to the tool and to then prove that the modifications have in fact improved the tool's performance. This is an extension to the approach made by Thomas et al. where they describe a sequence of trials.

This research is essentially the next logical step to the research that has been performed by Thomas et al. and it is anticipated that it will be of use to anyone who intends to move this research forward to a product development, or to anyone who wishes to extend the research into other diagram domains.

With these aims in mind, the following is the research question which is addressed by this project:

"Can the approach to the automated grading of imprecise diagrams as described by Thomas et al. (2006b) be extended over a larger sample of student answers, and using additional statistical metrics, so as to identify new deficiencies in the marking tool performance, and if so, can these deficiencies be corrected?"

Given these aims and research question, the following objectives were set for the research:

1) To define an experimental method and statistical analysis methods to evaluate the performance of the marking tool as provided. Primary concerns of these methods are:-

    a) To ensure that results are reproducible.

    b) The analysis methods should provide a measure of the how well the tool is performing (given the primary objective of the marking tool that each result should exactly match the moderated human mark).

    c) The statistical analysis methods should be capable of identifying deficiencies in the marking tool performance and should be useful for facilitating conclusions.

    d) The statistical analysis methods should have adequate sensitivity so that changes in the marking tool data are easily recognized.

2) To perform the experimental method on the 2 corpora of diagrams and use the statistical methods as defined in objective 1) to define and conclude on the marking tool performance.

3) Investigate potential of Bland-Altman methods to partially or completely fulfil the statistical analysis methods objectives as defined in 1).

4) To use the results of objectives 2) and 3), to design and conduct controlled experiments to determine some of the major root causes for some of the discrepancies between human and computer marked diagram answers.

5) To improve the marking tool performance by designing and implementing modifications to the marking tool. Performance of the modified marking tool is to be evaluated to confirm the improvement achieved and conclusions will be drawn.

1.6 Contribution to knowledge

It is anticipated that this project will result in a contribution to knowledge in the following areas.

1.6.1 Data Analysis

The Bland-Altman method is a method that is widely used within medical statistics for comparing 2 measurement methods. I believe that this method may provide an improvement over the statistical correlation measures that have been used by Thomas et al. (2006). Given this, I plan to investigate the application of this method for the comparison of human and automated marks. I will document the advantages and disadvantages of this method in the context of automated marking tools and will form conclusions regarding it's applicability within the automated marking tool context.

The audience for this knowledge would be developers of automated marking tools.

1.6.2 Understanding of the Imprecise Diagram Problem Domain

The investigations of Thomas et al. (2006) have demonstrated that as testing of the automated marking tool is extended, new and unexpected issues are discovered which lead to discrepancies between the mark awarded by the automated tool and the human mark.

The discovery and analysis of these issues will contribute to knowledge and insight into the domains of imprecise diagrams and automated marking. The audience for this knowledge would be developers of automated marking tools and other researchers working on the processing of imprecise diagrams.

1.6.3 Design Solutions for the Improvement of Automated Marking

One of the main objectives of this project is to design, implement and test improvements to the tool based on the discrepancies found between tool and human mark. So a natural result of this project will be knowledge and insights into design solutions for the processing of imprecise diagrams. The audience for this knowledge would be developers of automated marking tools and other researchers working on the processing of imprecise diagrams.

1.7 Overview of this dissertation

Chapter 2 of this dissertation contains a literature review. Within this chapter I have reviewed the published literature on automated marking tools with particular emphasis on the tool from the Open University research team.

Within Chapter 3 the research methods used within the project are discussed. These include the methods for marking tool test and performance analysis, and root cause analysis. Chapter 4 contains the test procedure used for testing the marking tool.

Chapters 5 and 6 contain the marking tool initial test results together with the root cause analysis of marking tool detractors.

In Chapter 7, modifications to the marking tool design are presented together with the results of the improvements.

Chapter 8 describes test results after the contents of Corpus A were extended.

Chapter 9 concludes this dissertation.

Chapter 2 Literature Review

I have grouped the body of knowledge into four categories for the purpose of this review. The first of these contains published material that relates to diagram marking tool implementations other than the Open University's tool. The second category contains the published material that directly relates to the diagram marking tool which has been implemented by the Open University's team and is the subject of this dissertation. The third category contains material of a general nature on natural language processing. The fourth contains general material on diagram theory and computing with diagrams. Within the last 2 categories I have selected the material that appears to be most useful to this project.

2.1 Survey of Automated Diagram Marking Systems

Within this section I have reported on the key findings from my literature search on other attempts that have been documented to automate the marking of student diagrams.

The first of these systems was reported by Hoggarth et al. (1998). Hoggarth et al. describe a system whereby a computer aided learning (CAL) system is embedded within a computer aided software engineering (CASE) tool. The system described was somewhat limited in it's usefulness to this project in that:

a) Use of CASE tool enforces drawing conventions rather than allowing a free form diagram.

b) The tool required the student to manually match symbols on the model answer to those of his own.

c) The tool only provided guidance feedback to the student highlighting the differences between the student and model answer. There is no attempt to grade the student's answer.

The second system that I reviewed is one that was developed by a team from Nottingham University. This tool appears to be the result of a long term project and its name has changed several times. The tool has been referred by the names of Ceilidh and CourseMaster previously, and its current name is CourseMarker.

This tool was originally developed with the primary function of teaching programming skills, and of particular interest was its ability in its early life to automatically grade student programming exercises. Higgins et al. (2001) described the CourseMarker system in some detail.

Both Tsintsifas (2002) and Higgins et al. (2002) described the extension of the CourseMaster system to manage and assess diagram answers. Key points are as follows:

a) The system is capable of presenting and assessing questions on logic design, flowchart design and Object-Oriented design

b) For logic design marking, a circuit simulator is used. This has to be preconfigured with test input data and expected output results.

c) For flowchart design marking, the marking module converts the flowchart into a program, and this is run using preconfigured test input data. Results are compared with the preconfigured expected output results.

d) For Object-oriented design marking, the student answer is inspected for predefined features such as the use of specific relationships between predefined classes

Higgins et al. (2005) provided an overview and status of CourseMarker but contains very little additional knowledge.

The following are my own observations on the information published on the marking of diagrams by CourseMarker:

1) The mechanisms for marking Logic design and Flowchart design diagrams are functional "Fit for Purpose" mechanisms. The student answers are modelled to evaluate if the process that is diagrammatically described gives the desired result. This approach has an advantage in that the course assessor doesn't need to present all alternative solutions to the marking tool. However, there are some disadvantages in that …

   a) The diagram input probably needs to be a precise input, rather than giving the student the opportunity for imprecision within the diagram.

   b) It's unlikely that there is any opportunity for the tool to offer partial marks for nearly correct answers.

   c) Each new diagram domain needs a new diagram processing method (rather than just needing a new diagram component language library).

   d) Many diagram domains do not have any possibility of a "Fit for Purpose" test, and for other domains such a test would be so complex that the implementation

would be unviable. So the applicability of this method to the general problem of marking diagrams is limited.

2) The marking mechanism for Object Oriented design diagrams is conceptually completely different from the other mechanisms and appears to be more akin to the system used in the Open University marking tool. However, the lack of explanation of the mechanism gives little useful insight for this project. In particular, there is no explanation of how the marking tool manages nearly correct solutions, synonyms used in Class or relationship names, or student freedom to produce imprecise diagrams that still have meaning.

3) No data is published to correlate human and marking tool results for the CourseMarker tool.

The final diagram marking tool that I have reviewed is described by Batmaz et al. (2006). Batmaz et al. describes a semi-automated marking tool. The primary functionality of this tool is to group student answers so that those with identical segments are in the same group. The marking tutor then only has to mark each group of answers once rather than marking student answers individually, hence saving marking time. The level of detail provided and the limited functionality of this tool led me to believe that the work of Batmaz et al. has limited usefulness to this project.

2.2 The Open University Diagram Marking Tool

In this section I've reviewed the publications by the Open University team on their marking tool which is the subject of this dissertation. I've taken the publications roughly in chronological order and I've highlighted the new insight into the marking tool that each publication provides.

The first publication on the Open University diagram marking tool was from Thomas (2004a). Thomas described his previous work on a marking tool for textual answers and explained that the work on the diagram marking tool was natural progression both in terms of addressing a limitation of text marking tools, and in terms of correspondences between diagrammatic and sentential forms. A feasibility study into the automated marking of diagrams was described including initial results. At this stage, the feasibility study was limited to a diagram requiring only text boxes and connecting lines. Of particular interest are the description of the use of a constraint multiset grammar (CMG) to represent and parse diagrams, and the outline description of the marking mechanism. I have given further explanation on CMG in the later section on the theory of diagrams.

Smith et al. (2004) focused on the imprecise nature of student diagram answers but offered little additional insight into the marking tool.

Waugh et al. (2004) described the concept of extending the marking tool to assess E-R diagrams. The five stage architecture to interpret diagrams was introduced but the descriptions of the stages were incomplete. A useful description of the marking mechanism using minimal meaningful units (MMUs) was included.

Thomas (2004c) focused on the practical aspects of using a drawing tool within an examination and gives no insight into the marking mechanisms.

Thomas (2004b) reported on the first significant field testing of the marking tool. A question requiring an E-R diagram was given to 26 students. Comparisons were made between human and marking tool results for the student's answers. Thomas showed a scatter plot of the results together with descriptive and correlation statistics for the 2 sets

of results. Insight was given into corrections needed for human marking error and application of marking rules within the tool.

Thomas et al. (2005) gave a summary of the same experiment and results already published in Thomas (2004b), but also gave detail of a second field trial involving a different, and more complex E-R diagram question. Results were reported as both pleasing and satisfactory. Disappointingly, the results were presented only in terms of mean, standard deviation, and Pearson and Spearman correlation statistics, and no scatter plots were presented. So the comparative performance for the marking tool across the full range of marks is not easily understood from the data presented.

Thomas et al. (2006a) introduced a number of key concepts. The first of these is the concept of equivalent diagrams. An example of this is the decomposition of a many to many relationship between 2 entities. The second concept introduced is the use of sub diagrams for automated grading. The third useful concept is that of a pattern, which the authors described as the general shape of a diagram. The authors proposed that patterns have 2 uses. These are to analyse diagrams into a set of sub-diagrams, and to synthesize a diagram from a set of sub diagrams.

Within section 4, Thomas et al. gave examples of patterns and showed how patterns can be matched within a sample ER-diagram.

Section 6 is important as Thomas et al. described in some detail the problem of inexact matching. This is an extension of the argument on the problems of imprecise diagrams that was previously introduced. Here, Thomas et al. described their approach to finding the best match and for measuring how close the match is. They went on to describe their method for measuring similarity.

Thomas et al. (2006b) presented a useful review of the diagram reasoning literature, their approach to diagram marking, and an overview of the trials and results achieved. Of particular interest is the discussion of the synonym problem that occurs when students are able to choose their own naming conventions for entities and relationships within their E-R diagram answers.

Thomas et al. (2007) expand their discussion on the problem of synonyms and also introduce the problem of marking E-R diagrams containing supertype-subtype relationships.

Together, these publications give significant insight into the Open University's diagram marking tool project. Important and detailed information was given about the architecture and mechanisms used within the marking tool to interpret imprecise student diagram answers, compare to a model answer, and to assign a mark for the student diagram. Additionally, a substantial amount of test data has been presented, comparing human marks with those from the marking tool. Results of investigations into marking tool discrepancies give insight into the practical problems encountered.

As minor criticism, one aspect that is not addressed within these publications is the justification for the architectures and mechanisms chosen or any discussion of the alternatives that might be available.

I was unable to find comparable documentary evidence on any other diagram marking tool either in terms of equivalent evidence of performance, detail of the marking mechanisms or of the problems encountered and overcome. This leads me to the conclusion that the Open University tool is probably the most advanced diagram marking tool project at this time.

2.3 Natural Language Processing

I've reviewed two titles on the subject of natural language processing. These provide no information of significant relevance to the diagram domain. However, they do go some way towards providing insight, and potential solutions to the problems of recognition of entity and relationship names that are given by students within their E-R diagrams as text, and matching them to the names within the model answers.

Jurafsky and Martin (2000) provide a detailed and broad introduction to natural language and speech processing. Much of this book has little relevance to this project. However, there are several chapters that are potentially useful.

Chapter 3 of this text discusses word morphology in the English language. It describes a number of scenarios. For many words, the different morphological forms simply have different endings (such as map, maps, mapping and mapped). Other morphological formations are more complex (such as goose and geese, or butterfly and butterflies). The principles of Finite-State Morphological Parsing and Finite State Transducers are introduced. A description of the Porter Stemmer is given. This stemmer goes some way towards resolving morphology issues. An implementation of the Porter stemmer is used in the E-R Diagram Marking Tool.

Within chapter 5, Jurafsky and Martin discuss another potential problem for the recognition of entity and relationship names, which is spelling errors. A discussion is given on the nature of spelling errors and their patterns, including probabilities. The Bayesian (or noisy channel) approach to spelling correction is described. This uses the statistical probability information to determine the most probable correct word. As an example, 80% of spelling errors have just one character in error. So if "th" were typed, then there is a large probability that the intended word was either "the" or "to" as each

of these is reached with one character correction. It is much less likely that the intention was to type “there” which requires three character corrections.

Chapter 13 discusses one other set of potential problems for recognition of entity and relationship names. These are the problems of synonyms, hyponyms and hypernyms.

Jurafsky and Martin discuss how WordNet, a database of lexical relations, is structured and can be used to resolve such problems.

As with Jurafsky and Martin, Manning and Schutze (1999) take a broad approach to natural language processing. Again, the majority of this title has limited usefulness.

Since both Jurafsky and Martin, and Manning and Schutze are similar in nature, I have used the Jurafsky and Martin text as a reference and compared the Manning and Schutze volume against this.

Manning and Schutze section 1.4.1 provides a listing and description of lexical resources available, whereas Jurafsky and Martin only identified WordNet.

Chapter 3 provides an overview of linguistic concepts including word morphology. However, much of the chapter is devoted to sentence formation which is irrelevant for this project. The coverage by Jurafsky and Martin is far more useful.

Whilst Manning and Schutze give somewhat less detail than Jurafsky and Martin on stemming, they do mention an alternative to the Porter stemmer which is the Lovins stemmer.

Manning and Schutze do not provide any information on the subject of spelling errors. The noisy channel model is discussed but this is done for problems other than spelling

errors. Similarly, whilst the subject of synonyms is covered, it is not of use to this project as the coverage is with respect to disambiguation.

2.4 Theory of Diagrams, Diagram Processing and Computing with Diagrams

A useful overview of general diagram methodologies has been presented by Blostein (1996). Importantly, Blostein asserted that there was no criteria for choosing a diagram recognition framework or how best to represent and apply notational conventions. Blostein's overview included sections on diagram recognition processes, diagram recognition knowledge and frameworks for diagram recognition, all of which are relevant to this project.

A number of published works have been found that focus on visual grammars and the parsing of these.

Marriott et al. (1998) give a useful introduction and overview of visual language specification. The use of grammars as frameworks for the production of visual languages is described, and a survey of available grammars is given.

Marriott et al. describe the importance of a grammar in providing a specified syntax to be used for creation of a visual language. The syntax rules provided by the grammar define the production rules for complex graphical objects which are composed of simpler graphical objects. The general approach of Attributed Multiset Grammars is described, of which, Constraint Multiset Grammar (CMG) is a subset.

As stated by Marriott (1994), and Chok and Marriott (1995), "Constraint multiset grammars provide a general, high-level framework for the definition of visual languages".

The main properties of CMGs are:

- CMGs are structured as sets of simple graphical shapes or symbols (so called multisets). This significance of this is that unlike text, for diagrams the sequence of construction has no relevance.

- Each shape object has attributes associated with them (e.g. a line might have attributes of a start location and an end location).

- Relationships between different shape objects and between a shape object and text are explicit within a production (in other frameworks this can be implicit by proximity).

- A set of constraints are applied to the diagram production.

Thomas (2004a) described the usage of CMGs for the representation of a diagrams, and demonstrated how to use them to define the visual language for a process pipeline. Thomas included in his example the representation of the smallest meaningful unit of discourse for the language which is an association, consisting of a pair of boxes connected by an arrow. Thomas demonstrates the formation of a constraint using the CMG notation to specify the construct of an association being the smallest unit for the production.

Chok and Marriott (1995) provide useful additional description and examples of the CMG together with a discussion on the parsing of the grammar.

Likewise, Tanaka (2001) and Iizuka et al. (2001) extended the ideas of constraint multiset grammars, adding further detail on parser techniques and extending the

knowledge to a drawing editor application, but these references do not add any information that is important to this project.

Rekers and Schurr (1997) provided an alternative perspective, using layered graph grammars. This publication is of particular note as extensive E-R diagram illustrations are used.

Chapter 3 Research Methods

The nature of this research project requires the following activities to be performed:

1. To design an experiment to evaluate the marking tool performance using the available diagram corpora. Perform that experiment.

2. Use the results of that experiment to determine areas where the marking tool performance is deficient and the causes of that deficiency (root cause analysis).

3. To design and implement reasoned modifications to improve the marking tool performance.

4. Repeat the marking tool evaluation, as defined in a) and compare the marking tool results before and after modification.

These activities and the research methods used to perform them are detailed in the following sections.

3.1 Evaluation of the Marking Tool Performance

Three sub-activities were performed as part of this evaluation...

- To automate the marking tool testing and data recording.

- To define a test procedure in adequate detail so that results are accurate and reproducible (a controlled experiment), and to use that procedure to gather the initial raw data.

- To define a set of statistical summary data together with graphical methods, and

use them to define the marking tool performance and identify deficiencies.

3.1.1 Automated Testing and Data Recording

The marking tool was modified to automatically select the default testing conditions and to automatically record data. These modifications add to the integrity of the test as a controlled experiment, reducing opportunity for error.

The automatic data recording was done by writing results to a CSV file format. This format is compatible with the XLStatistics package (Carr, 2004) used for the statistical analysis.

3.1.2 Test Procedure

The purpose of the test procedure is to ensure that testing gives correct results and is reproducible (i.e. to ensure that the process of testing is a controlled experiment). My test procedure addressed 3 areas:

- Verification that results output to the CSV file matched those generated by the marking tool.
- Experimental precautions which addressed opportunities for error in the usage of the marking tool.
- The marking tool has 9 parameters which can be adjusted by the user. Each of these modifies some aspect of the marking tool performance. Taking each parameter in turn the optimum value for each parameter was determined and used as the default for testing.

3.1.3 Statistical Measures and Graphical Presentation

The objectives of the data graphical presentation and statistical analysis measures are as follows:

1. The descriptive statistics should provide a measurement of the how well the tool is performing (given the primary objective that each marking tool result should exactly match the moderated human mark).

2. The descriptive statistics and graphical methods should be capable of identifying deficiencies in the marking tool performance and should be useful for facilitating conclusions.

The measurements derived from the statistical data, together with the graphical representations will be used in 3 respects. Firstly, the tool will be used for 2 different corpora of diagrams and an ideal measurement will be able to indicate whether the marking tool performance is identical for the 2 corpora, or whether there is some difference in performance. Secondly, the graphical and statistical data will be used to identify deficiencies in the marking tool performance. Finally, the marking tool will be modified and its performance re-evaluated. An ideal set of measurements will be able to indicate whether the tool modifications have improved (or degraded) the marking performance and will quantify that improvement.

To achieve these requirements, the descriptive statistics and graphical methods must have adequate sensitivity so that changes in the marking tool performance are reflected by changes in the measured values and graphs.

3.1.3.1 Descriptive Statistics

Table 3.1 and Table 3.2 together form a complete set of the descriptive statistics which are to be used. The results quoted are those for Corpus A.

	Corpus A Marking Tool Mark	**Corpus A Human Mark**
Mean	3.2056	3.1421
Standard Deviation	1.4845	1.5486

Table 3.1: Comparison of human and marking tool results

Measure	**Result**
Number	197
Percent Correct	64% (126/197)
Pearson Correlation Coefficient	0.9599
Mean of Absolute Value of Deltas	0.236

Table 3.2: Descriptive statistics for marking tool performance (Corpus A)

No one measure has been found which completely describes the marking tool performance.

A comparison of the means of the human and marking tool marks gives an indication of the presence of systematic marking errors. Usefulness of this comparison as a measure of absolute marking tool performance is somewhat limited. It is easy to imagine that it is possible for the means of the human and marking tool results to match in situations where no individual student diagram has been marked correctly. Conversely, it is also easy to imagine a situation where the large proportion of diagrams are marked correctly, but a large difference between the means occurs due

to a small number of large errors. Comparison of the means of human and marking tool results is useful as an indicator of marking tool improvement after modification since, as the marking tool is improved the Human marks remain unchanged.

The percentage correct is an important measure as it's directly related to the primary objective of having all marking tool results match the human mark (100% correct). However, this measure has the major disadvantage that it gives no indication about the results that did not match. So, for instance, this measure has no sensitivity to small improvements to extreme results.

The Pearson correlation coefficient is useful as an indicator of performance. It has the advantage that all results influence the resulting coefficient. Pearson correlation coefficient could potentially be used to compare marking tool performance across different corpora. But it has the disadvantage that it measures the strength of the linear correlation and is not a measure of agreement. So, for example, the marking tool could always give a mark 2 greater than the human mark and this would give perfect correlation (Pearson Correlation Coefficient $r = 1$) even though there were no correct results.

The mean of the absolute value of deltas is calculated by taking the absolute value of the difference between the human mark and the marking tool mark, and then calculating the mean value of these. This measure was found to be useful for comparing marking tool performance before and after improvement as it has a higher sensitivity than the Pearson Correlation Coefficient.

It should also be noted that for the comparison of the performance of the marking tool when marking the 2 corpora, only the percentage marked correctly and the Pearson

Correlation Coefficient should be used. All other measures are dependent upon scaling. To give an example, if the Corpus A question had a total mark of 70 instead of 7, then all other measures would be multiplied by 10. Values for percentage marked correctly and Pearson Correlation Coefficient are unaffected by scaling and would be unchanged. It is only valid to use other measures for comparison of marking tool performance as the marking tool is modified for the same corpus of diagrams.

The Spearman rho and Kendall's tau coefficients have been used previously as a measure of the tool performance (Thomas et al., 2006b). However, these measures are most appropriately used where data ranking is the prime concern. For these experiments, data are available in interval form and I don't believe that the Spearman coefficient adds any value to the analysis. Therefore, the Spearman and Kendall coefficients are not being used as measures.

3.1.3.2 Graphical Representations

Graphical representations perform a number of functions as follows:

- Visual confirmation that data are uniformly distributed. A non-uniform distribution could significantly affect the Pearson correlation coefficient (StatSoft Inc., n.d.).
- Identification of trends in the data.
- Visual comparison of graphs from different corpora can identify differences in marking tool performance.

- Identification of data points with the greatest errors which are to be used for root cause analysis.

Three graphical representations were investigated. Firstly I used a scatter plot of the Human Mark plotted against the best marking tool mark. This was overlaid with a least squares fit line (Figure 3.1). I compared this to the second representation of the same scatter plot overlaid with a y = x straight line (Figure 3.2).

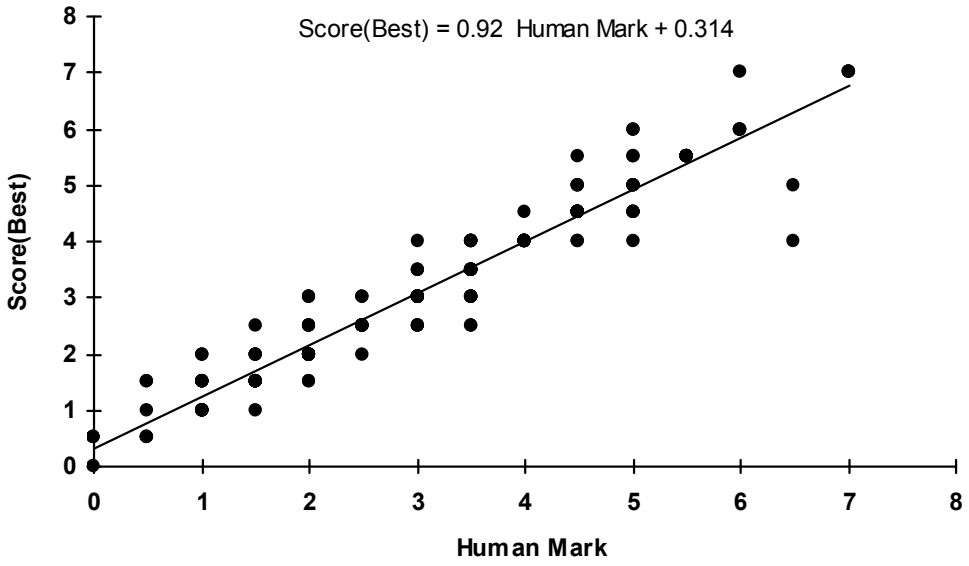

Figure 3.1: Scatter plot with least squares fit line for Corpus A results

I found that the least squares fit overlay confused the situation when looking for points with the greatest error. So this graphical representation is not going to be used.

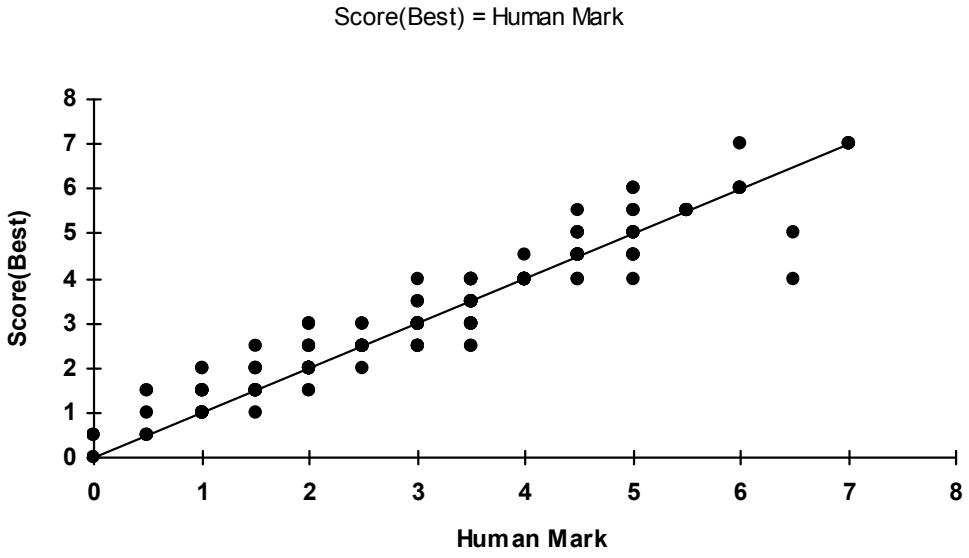

Figure 3.2: Scatter plot of best score from the marking tool vs human mark for Corpus A. y = x plot is overlaid.

The third graphical representation is discussed in the following section.

It should be noted that there is a serious limitation of the Scatter Plot. The marking tool and human mark data are in the form of discrete values as marks can only take values in increments of 0.5 marks. This leads to the issue that multiple data can occupy the same point on the plot and hence there are far fewer points on the scatter plot than there are data. This could have an influence on the visual judgement of data symmetry and uniformity of distribution.

3.1.3.3 Bland–Altman Methods

Kirkwood and Sterne (2003) provide detail of this method which is somewhat different from the Pearson and Kendall correlation tests used by Thomas et al. (2006). Dallal (n.d.) reviews the Bland-Altman method within the section of his web tutorial titled "Comparing Two Measurement Devices". In this review, Dallal describes 3

weaknesses of the linear correlation method and the reasons why the Bland-Altman method overcomes these.

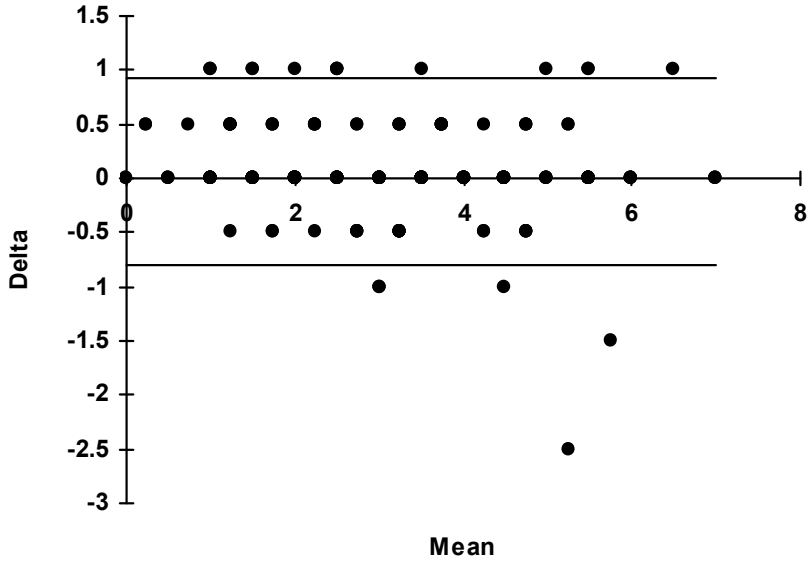

Figure 3.3: Bland-Altman plot of marking tool minus human mark (vertical axis) against mean of marking tool and human mark (horizontal axis) for Corpus A. The horizontal lines correspond to the 95% limits of agreement.

Figure 3.3 shows the Bland-Altman plot for Corpus A data.

The advantages of the Bland-Altman plots over scatter plots are as follows:

- The chart format is aligned with the primary objective of achieving identical results for the marking tool and human mark. Deviations from this objective result in a plotted point that does not lie on the x-axis.

- The sensitivity of the y-axis is increased so it is easier to recognise data points which have the largest discrepancy (the y-axis range is determined by the largest

discrepancy and not, as is the case in the scatter plot, the full range of the human marks).

- The data is presented with respect to a horizontal line rather than a diagonal, again making it easier to determine the largest data discrepancies.

It should be noted that the Bland-Altman plots suffer from the same problem of plotting discrete data as that of the scatter plot in that more than one item of data may occupy the same point on the graph. The Bland-Altman plot can only produce the same number of data points as the scatter plot because the data points on the Bland-Altman plot are derived directly from the human and marking tool scores which are the x and y values on the scatter plot.

To supplement the Bland-Altman plots, the 95% limits of agreement are calculated. These data can be used in addition to the other descriptive statistics described above. The 95% limits of agreement for the Corpus A data are shown in Table 3.3.

Parameter	**Value (Corpus A)**
Mean of Marking Tool Mark minus Human Mark	0.0635
Standard Deviation of Marking Tool mark minus Human Mark	0.4342
Upper 95% Limit of Agreement (mean + 2 standard deviations)	0.9319
Lower 95% Limit of Agreement (mean - 2 standard deviations)	-0.8049

Table 3.3: 95% limits of agreement for Corpus A results

These parameters are useful in that they indicate the data trends. So it is easy to see that the marking tool has a slight positive bias for Corpus A. Also, they might be useful for comparison of the marking tool performance before and after modification. I would expect to be able to use these parameters as a guide to the nature of the effect

of any modification. However, they appear to be of little use as an absolute measure of performance. For that purpose I would still expect to use the measures as indicated in Table 3.1 and Table 3.2.

3.2 Root Cause Analysis

Diagrams where the marking tool result shows significant disagreement from the human mark will be selected according to the magnitude of the marking error, and these will be used for root cause analysis.

Three basic methods are expected to yield root causes for marking tool error.

The first method is a form of controlled experiment. Each diagram will be manipulated in a controlled manner so as to ascertain the features of that diagram which cause the marking tool to produce an incorrect result. For instance, if it is suspected that the relationship names were wrongly interpreted by the tool, these names can be corrected and the diagram marked again to confirm the effect of the relationship name.

The second method is a logical investigation of the marking tool behaviour when marking the diagram under investigation. Detailed analysis can be made of object instance states during the marking process. These can then be compared to those of similar diagrams which have been marked correctly. This should give insight into the way that the marking tool behaviour has to be changed to give a correct result. This method is somewhat similar to the standard practice of code debugging by printing the status of variables as the code runs.

The third method is to manipulate the marking tool parameters and retest the diagram

that has been incorrectly marked. This will ascertain if the marking tool processes as they exist are capable of producing a correct answer. This method will be used selectively where indicated by the results of the first 2 methods.

Two generic analysis methods have been considered and rejected at this time. Pareto analysis has been considered for the root cause analysis as a method of prioritising the root causes with the greatest overall impact (The Open University, 1995). This method is rejected as the effort needed to analyse root cause of a significant sample of diagrams is considered to be too great. The second rejected method is that of using Ishikawa diagrams as a method of starting with an effect and decomposing this into the primary causes that must be tackled (The Open University, 1995). I believe that for this project that the network of causes and effects will not have adequate complexity to justify this method.

3.3 Marking Tool Modifications and Testing

When the primary root causes of marking tool error have been identified, these have to be translated into improvements in the marking tool.

To achieve this, a prototype needs to be produced. The stages to produce a prototype will be design, implementation and testing. It is likely that these 3 stages will be performed iteratively in order to produce a satisfactory prototype.

The design and implementation stages will focus on the project objectives of demonstrating that an improved marking performance can be achieved, rather than on any aesthetic concerns such as usability.

Testing will be performed using the methods described above and data will be

presented as a comparison between the original data and the data after modification.

Chapter 4 Test Procedure

This section describes in detail the experiment and results which were outlined in section 3.1.2.

4.1 Verification of Marking Tool Output

Verification of marking tool output identified 2 bugs within the marking tool code. The first of these caused the marking tool to fail to return the best scores where the marking was performed using 2 solution diagrams. The second of these prevented the parameter "Relation Name Weight" from being altered using the dialogue box within the tool. The code fixes to correct for these is documented in Appendix A.

4.2 Experimental Precautions

As one would expect, the automated marking tool required very little human intervention to operate. However, 2 areas of concern were identified.

Firstly, the marking tool has a number of parameters that affect marking performance. After determining the values for these parameters, the same values were used for all testing. To minimise the possibility of performing tests using the incorrect parameter settings I changed the marking tool code so that the parameters are set correctly when the application is started.

Secondly, root cause analysis required some editing of the student answer diagrams which could lead to a situation where the primary data is compromised. To eliminate the possibility of inadvertent editing of diagrams (student answers or solution diagrams), diagrams were stored within a read only directory, and copied from there

to a second read-write directory when marking was to be performed.

4.3 Determining the Values of the Marking Tool Parameters

The marking tool has a number of parameters which have their values set by the application user prior to performing marking. The setting of these parameters allows the user some choice as to whether certain subroutines are invoked during the marking operation, and allows the user some control over certain variable limits that are used during marking process by the tool. Table 4.1 lists the parameter names, the range of values that may be used, and the optimum value that was used for testing prior to modification of the marking tool.

Parameter	Range	Optimum Value
Plausible Matching	False / True	True
Synonym Check	False / True	True
Stemming Enabled	False / True	True
Assoc-Adorn Weight	0 to 1	0.6
Degree-Participation Weight	0 to 1	0.5
Relation Name Weight	0 to 1	0.1
Similarity Threshold	0 to 100	60
Plausible Weight	0 to 100	50
Closeness Threshold	0 to 1	0.9

Table 4.1: Marking tool parameter names, range of values and optimum values

In order to determine the optimum setting for these parameters, each was taken in turn and it's effect on the marking tool performance for the 2 corpora were measured. For parameters where only true or false values are allowed, only 4 tests were needed (2 parameter values for 2 corpora). Where the parameter value requires a number, the marking tool results for each of the corpora were measured over the entire allowable

range of parameter values, with the parameter value being incremented by 10% of the allowable range.

Test (Corpus A)	Mean	Correlation Coefficient
Human Mark	3.1421	
Plausible Weight = 30	3.2157	0.9585
Plausible Weight = 40	3.2157	0.9585
Plausible Weight = 50	3.2056	0.9599
Plausible Weight = 60	3.1624	0.9517
Plausible Weight = 70	3.1345	0.9499
Test (Corpus B)	**Mean**	**Correlation Coefficient**
Human Mark	2.4667	
Plausible Weight = 30	2.1167	0.9103
Plausible Weight = 40	2.1167	0.9103
Plausible Weight = 50	2.0833	0.9058
Plausible Weight = 60	1.9667	0.8875
Plausible Weight = 70	1.8500	0.8566

Table 4.2: Results for the automated marking as the Plausible Weight parameter is changed

Table 4.2 illustrates this process by showing how the average and correlation coefficient of the Corpus A and Corpus B results change with the value of the Plausible Weight parameter. In this instance the optimum value for the parameter was chosen to be 0.5. As can be seen, this value is a compromise between the Corpus A and Corpus B performance. For Corpus B, value of 40 for Plausible Weight would have yielded better results with a better correlation coefficient and a mean value closer to that of the human mark. But this value would have degraded the marking tool results for Corpus A, giving a worse correlation coefficient and a mean value further from the human mark.

It should be noted that this method assumes that the effect of each parameter on the marking tool performance is independent of the others. This was not validated and so I would consider this method of ascertaining parameter values to be an incomplete

assessment.

4.4 Summary

The development of the test procedure focussed on 3 areas:

1. Validation of the correctness of automated output from the marking tool software.

2. Controlling and documenting those parameters affecting the output.

3. Prevention of errors due to inadvertent alteration.

The establishment of the test procedure accomplished the objective of ensuring that results were accurate and repeatable.

Chapter 5 Test Results Prior to Marking Tool Modification

Tests were performed on the 2 corpora using the methods described in the previous chapters. Results for Corpus A are documented in Appendix B, and results for Corpus B are documented in Appendix C.

The question relating to Corpus A had 2 alternative solution diagram answers. In this situation, the marking tool scores the student answer using each answer diagram independently, and then selects the best score out of the 2. Hence, the Corpus A results have 3 columns containing scores, the Score(0) column contains the result of marking using the first answer diagram, the Score(1) column using the second answer diagram, and the Score(Best) column contains the best score out of the 2. It is the Score(Best) that is used as the marking tool's score for the diagram.

5.1 Descriptive Statistical Results

Table 5.1 shows the mean and standard deviation of the marking tool and human marks for the 2 corpora.

	Corpus A Marking Tool	**Corpus A Human Mark**	**Corpus B Marking Tool**	**Corpus B Human Mark**
Mean	3.2056	3.1421	2.0833	2.4667
Standard Deviation	1.4845	1.5486	1.1071	1.1290

Table 5.1: Comparison of the mean and standard deviation of the marking tool results and the human mark

For both corpora the marking tool marks have a smaller standard deviation than that of the human marks. If the opposite were true then I would have concluded that the

marking tool had introduced some random error element to the marking and a marking tool deficiency would have been the cause. This is not the case. The greater spread of human marks could be caused by a number of factors and may not be significant.

Table 5.2 contains the results for the percentage of student answer diagrams that were marked correctly, together with the Pearson Correlation Coefficients and the mean of the absolute value of deltas for the Corpus A and Corpus B diagrams.

	Corpus A	**Corpus B**
Percentage Marked Correctly (number marked correctly / number marked)	64% (126/197)	30% (9/30)
Pearson Correlation Coefficient r	0.9599	0.9058
Mean of Absolute Value of Deltas	0.236	0.483

Table 5.2: Percentage marked correctly and Pearson Correlation Coefficient for the corpora

Within this table the differences in marking tool performance between Corpus A and Corpus B are evident. In particular, the results for percentage of diagrams marked correctly demonstrate a very marked difference.

5.2 Graphical Results

Figure 5.1 is a scatter plot of the Corpus A data with the marking tool score on the y axis and the human score on the x-axis. Figure 5.2 is the same data plotted using the Bland–Altman method.

The data appears to contain 2 significant outliers at positions (5.25, -2.5) and (5.75, -1.5). If these are ignored then the data appears to be weighted towards the

positive Delta value (as there are many more points on the Delta = +1 line than there are along the Delta = -1 line). There appears to be no other significant skew or any non-uniformity that would affect the correlation coefficient.

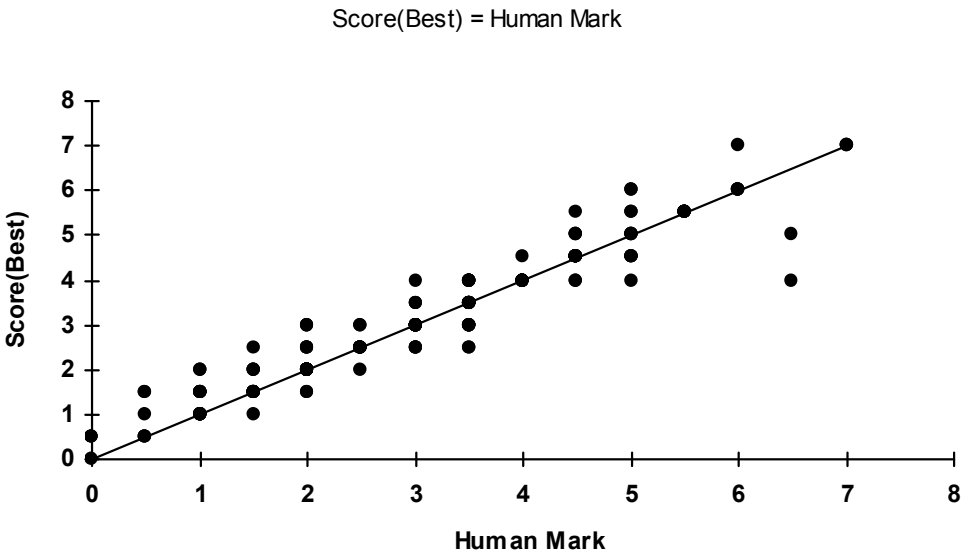

Figure 5.1: Corpus A data scatter plot with x = y overlaid

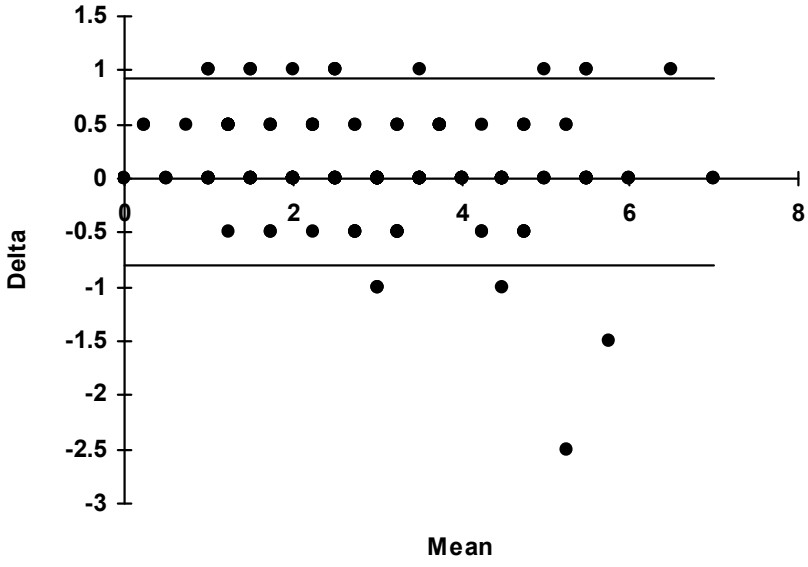

Figure 5.2: Bland-Altman plot for Corpus A data with 95% confidence levels overlaid

Measure for Score(Best) – Human Mark	Result
Mean	0.0635
Standard Deviation	0.4342
Upper 95% Confidence Level	0.9319
Lower 95% Confidence Level	-0.8049

Table 5.3: Descriptive statistics for the Bland-Altman plot for Corpus A data

Table 5.3 contains the data used to calculate the 95% confidence levels for the Corpus A Bland–Altman plot.

Figure 5.3 is a scatter plot of the Corpus B data. Figure 5.4 is the same data plotted using the Bland–Altman method. The data appears to be weighted towards the negative Delta value (as there are many more points below the Delta = 0 line than above it). There appears to be no data outliers, no other significant skew or any non-uniformity that would affect the correlation coefficient.

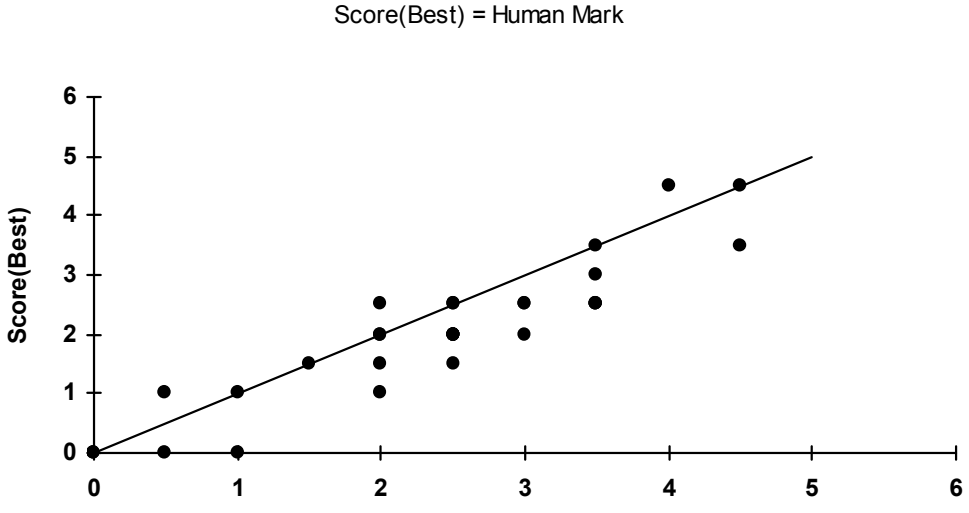

Figure 5.3: Corpus B data scatter plot with x = y overlaid

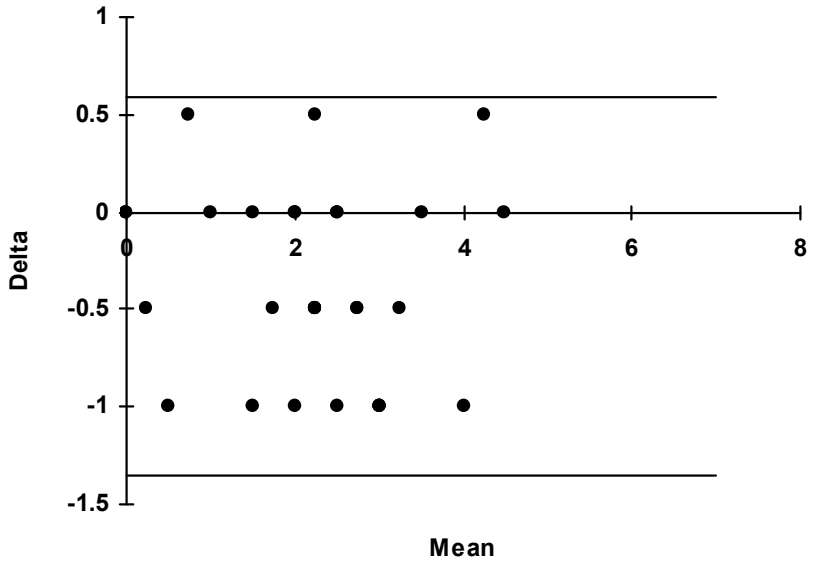

Figure 5.4: Bland-Altman plot for Corpus B data with 95% confidence levels overlaid

Measure for Score(Best) – Human Mark	Result
Mean	-0.3833
Standard Deviation	0.4857
Upper 95% Confidence Level	0.5881
Lower 95% Confidence Level	-1.3547

Table 5.4: Descriptive statistics for the Bland-Altman plot for Corpus B data

Table 5.4 contains the data used to calculate the 95% confidence levels for the Corpus B Bland–Altman plot.

5.3 Observations and Conclusions

The data shows that the marking tool performs very differently when marking the 2 corpora of data. The strongest evidence for this is the difference between the

percentage of student answer diagrams marked correctly (64% for Corpus A and only 30% for Corpus B). Additionally, the graphs show that the tendency for Corpus A diagrams is for the marking tool to provide a slightly high score, whereas, for Corpus B there is a strong tendency for the marking tool to provide a low score.

These observations were significant for the project as it moved to the root cause analysis stage as they indicate that different error mechanisms exist within the marking tool for the different corpora of data. In particular, the answer diagram for Corpus B contains subtype entities whereas the answer diagrams for Corpus A do not. So it was anticipated that the marking mechanism for subtypes was systematically giving a low score.

Using the Bland–Altman plots as a guide, it was decided that the root cause analysis should proceed using those student answer diagrams where the difference between the human score and the marking tool score is greater than or equal to 1. The value of 1 was chosen for the following reasons. Firstly, this project was time constrained and there was never any intention to analyse all diagrams with an incorrect mark. Choosing diagrams with a marking error of 1 or greater provided a systematic method of containing the sample for failure analysis to a manageable size. Secondly, choosing diagrams with the biggest marking error maximised the chances of identifying those problems causing large marking errors. Thirdly, the diagrams with the biggest marking error were most likely to be the easiest to identify a definitive error mechanism and are less likely to have human marking error as the only cause of discrepancy.

Chapter 6 Root Cause Analysis

6.1 Methodology

Data points were identified where the difference between the human mark and the making tool result was 1 mark or greater. The data points were then matched with student answer diagrams giving a list of diagrams on which to perform root cause analysis.

Marking tool output data was examined to obtain detail of the marks awarded for each relationship.

A table was drawn up for each diagram showing the attributes of the model answer which contribute towards the total mark (relationships and subtypes). The student answer diagram was inspected and the equivalent attribute on the student diagram was identified and added to the table (if any was present). The marking tool mark for each scoring attribute was added to the table. I made an assessment of the expected human mark for each attribute and added this to the table.

My assessment of the mark was based on the following understanding of the marking schemes. For the Corpus A marking scheme, 1 mark was available for each of the 7 relationships in the specimen solution that were drawn correctly by the student. Partial marks could be awarded for each relationship and the 1 mark was subdivided into 0.5 marks for correct participation conditions and 0.5 marks for correct identification of the degree of the relationship.

For Corpus B the same rules apply for the 4 relationships in the specimen solution. Additionally, 0.5 marks were awarded for each of the 4 subtypes in the specimen

solution that the student had correctly drawn.

It was clear during this process that the marking schemes were not a complete specification and could not cover all situations, nor did they describe fully describe the thought process of an experienced professional marker.

Scoring Attribute	Attribute Type	Match on Student Answer	Marking Tool Mark	Assessed Mark
Receives	Relationship	Attends	0	0
Gives	Relationship	Delivers	1.0	1.0
Prepares	Relationship	Prepared	1.0	1.0
BelongsTo	Relationship	Consists of	0.5	0.5
Presents	Relationship	(no match)	0	0
Has	Relationship	Requires	0	1
Is	Relationship	(no match)	0	0
Total			2.5	3.5

Table 6.1: Table of scoring attributes for diagram 10024 from Corpus A

Scoring Attribute	Attribute Type	Match on Student Answer	Marking Tool Mark	Assessed Mark
Attends	Relationship	Attends	1.0	1.0
AttendsCreche	Relationship	(no match)	0	0
Provides	Relationship	Runs	0	1.0
HeldAt	Relationship	Hosts	1.0	1.0
SlotIn	Relationship	(no match)	0	0
Parent	Subtype	(no match)	0	0
Child	Subtype	Child	0.5	0.5
External Event	Subtype	GoodStart	0.5	0.5
CoreService	Subtype	Other	0.5	0.5
Total			3.5	4.5

Table 6.2: Table of scoring attributes for diagram 10026 from Corpus B

Table 6.1 is an example of using this method for diagram 10024 from Corpus A.

Similarly, Table 6.2 is an example for diagram 10026 from Corpus B

The following steps were taken as part of root cause analysis:-

1. My assessed mark (column 5 total) was compared with the human mark. Diagrams were reassessed to try to resolve any discrepancies. Any unresolved discrepancy was logged as human error. It should be remembered that the human error category will include errors made by myself as well as those made by the original human marker. No differentiation was made between these two.

2. For some diagrams, the assessment / reassessment process revealed areas that were difficult for me to mark due features on the diagrams for which I had no understanding of the marking rules. For these diagrams the unresolved discrepancy was logged as indeterminate cause. A good example of this is diagram 24 from Corpus B. On this diagram none of the relationships are named. It is clear that the human marker must have awarded some, but not full marks for unnamed relationships.

3. Discrepancies between Marking Tool Mark (column 4) and Assessed Mark (column 5) were investigated one by one. Causes of errors were recorded and categorized. Where possible, positive validation was sought. This was achieved by manual modification of the student answer diagram and remarking the diagram in order to find the change in marking tool response. For instance, where the cause of discrepancy was suspected to be due to name recognition, the suspect names were changed on the diagram, one at a time, and the diagram was remarked after each change to measure the effect of the change.

When performing step 1, many diagrams were found to have human error. Many of these will of course be caused by my lack of experience or incomplete understanding of marking practice. A feature of the process that I adopted was that for some diagrams, a portion of the difference between marking tool and human mark was attributed to human error and another portion was attributed to another reason. So even in cases where I couldn't entirely match the original human mark, I did still sometimes retrieve useful data about marking tool deficiencies.

6.2 Results and Analysis

Results of the root cause analysis and categorisation are shown within Table 6.3. Note that a diagram reference notation has been adopted so that the first letter represents the Corpus containing the diagram, and the following digits represent the last digits of the file name.

Diagram	Human Error	Tool Incorrect Relationship Matching	Indeterminate	Entity Name Recognition	Relationship Name Recognition	m:n Relationship Decomposition
A1				◆		
A3			◆			
A24				◆		
A75			◆			
A198					◆	
A14	◆					
A34	◆					
A51		◆				
A84		◆				
A91	◆	◆				
A94		◆				
A108	◆					
A110	◆					
A118			◆			
A135	◆	◆				
A136	◆					
A152		◆				
A168	◆					
B4	◆					
B12	◆					
B15			◆			
B19		◆				
B24			◆			◆
B25	◆	◆				
B26	◆	◆				
B30			◆	◆		
Total Incidence	12	9	6	3	1	1

Table 6.3: Incidence and categorisation of marking discrepancies

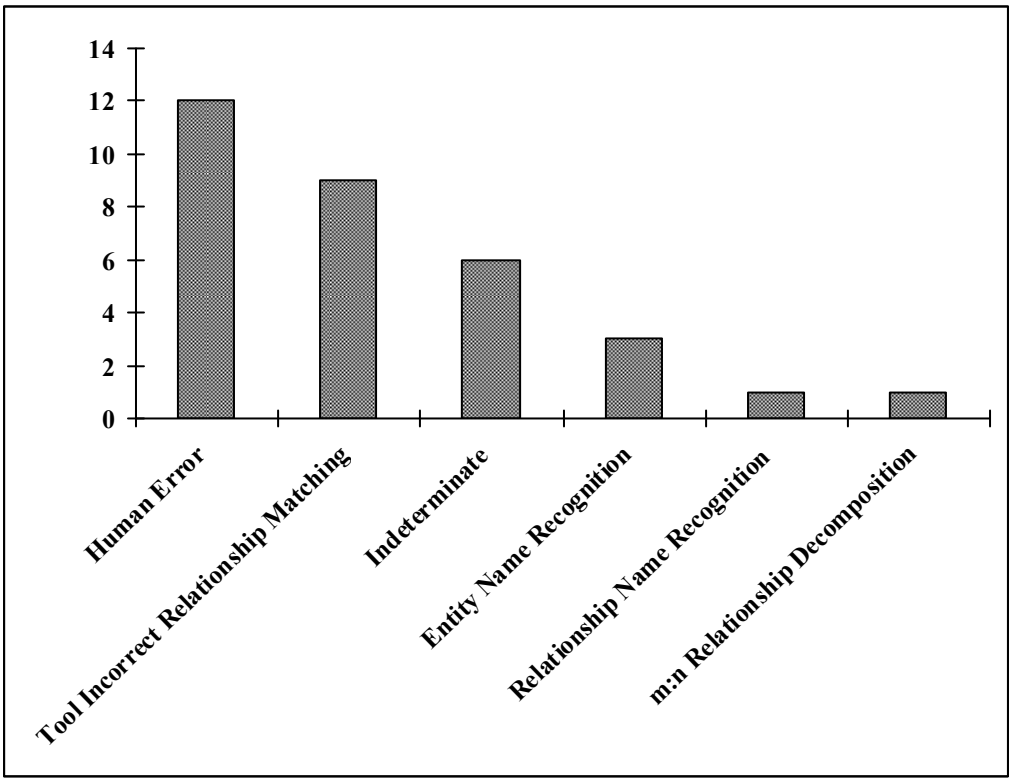

Figure 6.1: Incidence of root causes

Figure 6.1 shows a plot of the incidence of each cause. This analysis looks very much like a Pareto chart, however, I haven't referred to this as a Pareto analysis for 2 reasons:

1. The data samples are not random. They were pre-selected by choosing the diagrams where the marking tool had the greatest marking error.

2. The data comes from the marking of 2 different diagram corpora which are known to suffer errors due to different root causes. The 2 corpora are not equally represented.

Given these factors, it would be wrong to claim that Figure 6.1 is a true indication of the incidence of error causes across the population within the 2 corpora, or that it

represents the expected outcome if testing were extended. An analysis using Corpus A alone for example, would have yielded a very different chart.

It is also worth noting that the above analysis could have been performed by assigning a magnitude to each error occurrence and plotting the magnitude rather than the number of incidents. This wasn't done as I felt that the usefulness of such an analysis would be limited. In particular:

1. The sample of answers analysed was not random, so a quantitative analysis could not be used to estimate the potential for improving the marking tool performance when marking the entire corpora.

2. As well as having magnitude, the effect of the error can be positive or negative. It's unclear how to manage positive and negative effects quantitatively (i.e. if an error caused one diagram to be marked 0.5 marks above the human mark, and another diagram to be marked 0.5 marks below the human mark, is the net effect of this error zero or one mark?)

6.3 Root Causes

This section describes in more detail the nature of the error causes and how the categorization was done.

6.3.1 Human Error and Indeterminate

By definition, root causes within these categories are out with the scope of this project and for this reason I have grouped them together.

Analysis of these errors was not performed to the same level of detail as those errors

which are attributable to the marking tool. However, some useful observations were made.

The assignment of "Indeterminate" as root cause was made when a diagram had features which are not fully covered by the known marking rules, where I judged it reasonable that the human marker had perhaps applied some degree of discretion, or where I could not understand how to correctly mark a diagram.

There was one notable problem. In some student answers, some or all relationships had no name. The standard requirement for E-R diagrams is that all relationships should be named. However, my observation of the human marking practice was that partial marks have been awarded for relationships which are unnamed but which otherwise would have been awarded a mark. If a specific and complete set of marking rules were available for these situations then there would be opportunity to improve the marking tool in this aspect.

The assignment of "Human Error" as root cause was made when unaccountable discrepancies between the marking tool and human marks existed and where the discrepancy didn't fall into the "Indeterminate" category (i.e. no features outside the known marking rules existed, no discretion on behalf of the marker appeared to be warranted, and the diagram was straightforward to mark).

It should be noted that it would be wrong to use this data to draw any conclusions about the accuracy of the human markers. My assessment was performed without full knowledge of the entire moderation and marking criteria, so it is quite possible that in many instances, my assessment of Human Error could be inaccurate. To re-iterate, this doesn't matter since both Indeterminate and Human Error root causes are outside

the scope of this project.

6.3.2 Tool Incorrect Relationship Matching

Two distinct issues were identified which were included in this category.

The first of these is that the marking tool incorrectly matches a relationship on the student answer with a relationship on the specimen solution and awards marks.

Figure 6.2 showing diagram 10051 from Corpus A (A51) is typical of this situation.

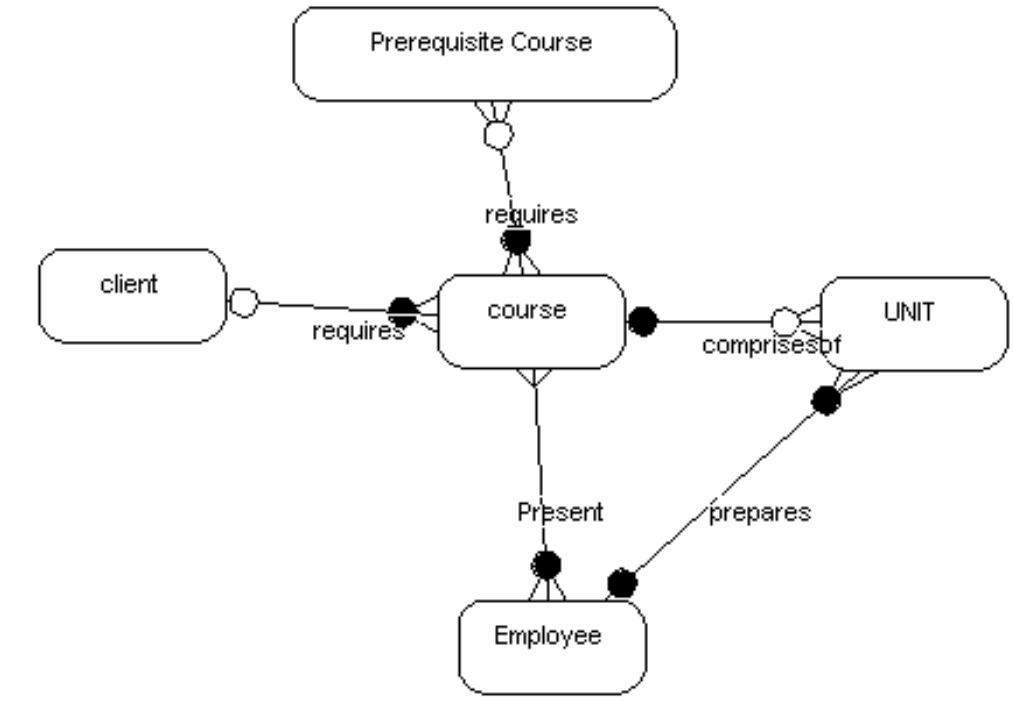

Figure 6.2: Student answer diagram 10051 from Corpus A

In this case the relationship "requires" between entities "client" and "course" was equated with the relationship "Receives" within the model answer and was awarded 1 mark by the marking tool (see Appendix for model answer diagrams). However,

within the model answer the relationship “Receives” connects entities “Client” and “Presentation”. It appears from the human mark that the marker didn’t equate these relationships and awarded no marks for “requires”.

This type of error was observed in the analysis of Corpus A diagrams only and in all cases caused the marking tool to provide a higher mark than that of the human marker.

The second distinct issue occurs in situations where the student’s answer is missing a subtype, if the model answer has a relationship connected to that subtype then it is impossible for the student’s answer to have a relationship connected to the missing subtype. In these cases, and where the student’s answer had an equivalent relationship to the super type, it was observed that the human marker would typically deduct appropriate marks for the missing subtype and award marks for the relationship as though it had been connected to the subtype. This approach seems to be appropriate as otherwise the student suffers a double penalty for one error. The marking tool cannot currently reproduce this marking strategy.

This type of error was observed in the analysis of Corpus B diagrams and in all cases caused the marking tool to provide a lower mark than that of the human marker

6.3.3 Tool Relationship and Entity Name Recognition

For a number of diagrams, it was possible to improve the marking tool mark by changing the name of an entity or relationship on the student’s answer. Table 6.4 documents these occurrences.

Diagram	Feature Type	Original Name	Original Mark	Revised Name	Revised Mark	Human Mark
A1	Entity	TrainingCourse	4.0	Course	6.5	6.5
A24	Entity	TrainingCourse	2.5	Course	3.5	3.5
A198	Relationship	IsAPrerequisiteOf	4.0	Is	5.0	5.0
B30	Entity	SupportedEvent	2.5	ExternalEvent	3.0	3.5

Table 6.4: Improvements achieved by revised naming

As indicated in the table, for 3 of the diagrams, the marking tool is capable of accurately marking them by changing just one name on the diagram. For the fourth diagram the mark is improved to within half a mark. Importantly, for diagram B30 the revised name eliminates the total error and the remaining half a mark error is attributed to other indeterminate causes.

6.3.4 m:n Relationship Decomposition

Within diagram 24 from corpus B (Figure 6.3), it was found that the many to many relationship "Attends" had been decomposed into the form of an entity named "Attendee" with appropriate relationships connecting from entity "Event" to entity "Person".

I judged that the human marker probably awarded 0.5 marks for this decomposition since the relationship degree is correct, the relationship naming is close to the model answer, but the participation conditions are not correct. However, the marking tool did not award any marks.

I found that this situation could be corrected by adding names to the relationships connecting the entity "Attendee" with "Person" and "Event".

I had earlier stated that it was a requirement of E-R diagrams that all relationships should be named. However, I believe that it is common practice not to name relationships when decomposing a many to many relationship into a new entity and two relationships. Usually, the new entity is named according to the name of the original many to many relationship.

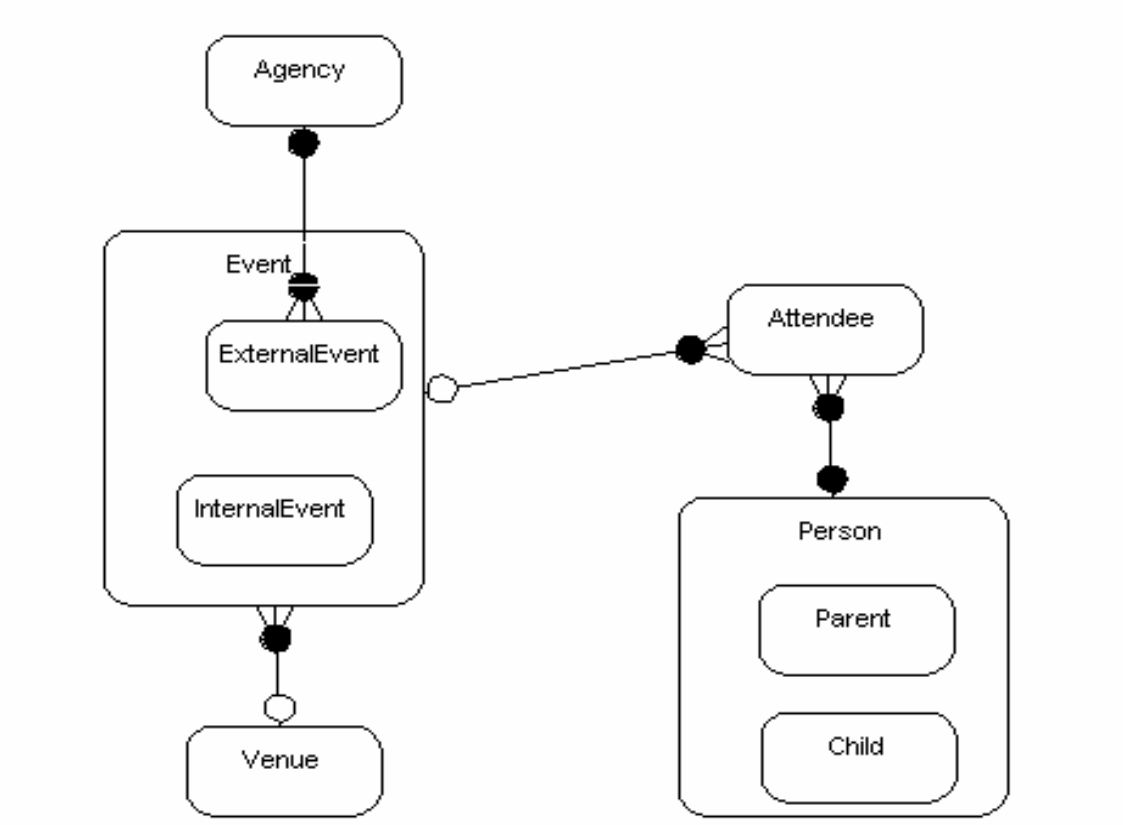

Figure 6.3: Student answer diagram 24 from Corpus B

Chapter 7 Marking Tool Improvements

This chapter describes which issues have been progressed from the root cause analysis towards a solution within the marking tool. For each issue that is addressed, the design and implementation of the marking tool modifications is described together with the marking tool results after modification.

7.1 Incorrect Relationship Matching for Corpus B Diagrams

This issue is the second of the distinct issues found which caused incorrect relationship matching by the marking tool. The issue occurs only in situations where the student answer is missing a subtype. Section 6.3.2 describes how this issue arises due to the marking tool's lack of capability to adopt the same marking strategy as the human marker in cases where a subtype is missing. Essentially, where a subtype is missing, the human marker will award marks for a relationship which should be connected to the missing subtype, if the relationship has been connected to the super type.

7.1.1 Design and Code Implementation to Fix

The following strategy was used within the design of the marking tool modification to fix this issue:

1. After the normal scoring process, missing subtypes are detected within the student answer diagram.

2. A copy of the object representing the specimen solution E-R diagram is created.

3. The missing subtypes are processed one by one. For each missing subtype, all relationships within the copy of the specimen solution are checked for connection to the missing subtype.

4. When discovered, the relationship is modified so that instead of connecting to the missing subtype, the relationship connects to the super type.

5. When all missing subtypes are processed, we are left with a modified copy of the specimen solution. The student answer is then re-scored using this modified copy. The highest out of the 2 scores is recorded.

This design strategy has the advantages that:

1. All existing marking tool code remains intact so there is no effect on the results for marking Corpus A diagrams.

2. It is easy to identify the original and new results (as both results are generated).

3. The code can be activated / deactivated as a user controlled option depending on whether this marking strategy is to be used for a question.

A copy of the implementation code is contained within the Appendix.

7.1.2 Test Results

Initial testing showed only marginal improvement. However, from examination of the marking tool output data it was recognised that a small change to the marking tool parameters might improve the performance. An investigation was made to test this

phenomenon.

Data was gathered to measure the effect of the Similarity Threshold parameter on the modified marking tool when used for testing Corpus B diagrams. This data is presented in Table 7.1.

Table 7.2 and Table 7.3 show the effects of changing the same parameter using the un-modified marking tool. As can be seen from these tables, the choice of value for Similarity Threshold is very much a compromise, with the marking tool performance for Corpus B degrading severely at values of 70 or above, and the marking performance for Corpus A diagrams degrading slightly at values below 60.

Similarity Threshold	Qty Diagrams Marked Correctly	Qty Diagrams Marked with Error 0.5	Qty Diagrams Marked with Error 1.0	Mean	Corr. Coefficient
30	11	15	4	2.25	0.9108
40	11	15	4	2.25	0.9108
50	11	15	4	2.25	0.9108
55	10	14	6	2.167	0.9065
60	10	14	6	2.167	0.9065
70	5	15	10	1.983	0.9051
Human	-	-	-	2.467	-

Table 7.1: Performance of marking tool including Substitute Marking method for Corpus B diagrams as Similarity Threshold is changed

Similarity Threshold	Qty Diagrams Marked Correctly	Qty Diagrams Marked with Error 0.5	Qty Diagrams Marked with Error 1.0	Mean	Corr. Coefficient
30	9	14	7	2.133	0.9013
40	9	14	7	2.133	0.9013
50	9	14	7	2.133	0.9013
60	9	13	8	2.083	0.9058
70	5	13	11	1.900	0.9024
Human	-	-	-	2.467	-

Table 7.2: Performance of unmodified marking tool for Corpus B diagrams as Similarity Threshold is changed

Similarity Threshold	Qty Diagrams Marked Correctly	Qty Diagrams Marked with Error 0.5	Qty Diagrams Marked with Error 1.0	Mean	Corr. Coefficient
30	124	51	20	3.234	0.9597
40	124	51	20	3.234	0.9597
50	124	51	20	3.234	0.9597
55	124	50	20	3.231	0.9598
60	126	53	16	3.206	0.9598
70	124	59	11	3.168	0.9606
Human	-	-	-	3.142	-

Table 7.3: Performance of unmodified marking tool for Corpus A diagrams as Similarity Threshold is changed

Comparing Table 7.1 and Table 7.2 it is clear that the modifications made to the marking tool have improved marking performance with Similarity Threshold set anywhere in the range from 30 to 60. It appears reasonable at this point to document results using a Similarity Threshold of 50. In order to get a true comparison, these are compared with the results of the un-modified marking tool using the same value.

In commenting on the parameter setting, Thomas (2008) advised me that the fundamental approach of the marker is to compute a measure of the similarity

between relationships (including subtypes) and entities and then determine which parts of a student answer best match parts of the specimen solution. The similarity measure uses a number of parameters, some of which are set by the user of the system (in a sense these represent features of the domain – the fact that E-R diagrams are being marked) and some which are found by experiment (and reflect the data). So it comes as no surprise to learn that the parameters for one corpus might be different to those for another corpus.

Table 7.4 shows which diagrams' marks were affected by this marking tool modification. Out of 6 marks affected, 1 mark was degraded (further away from the human mark) and 5 marks were improved. The modification had no impact on diagram B25 suggesting that further investigation and fine tuning of the modification may be required.

Table 7.5 and Table 7.6 contain a comparison of the descriptive statistics for the Corpus B data using both the unmodified and modified marking tools.

Diagram	Human Mark	Marking Tool Mark Prior to Modification	Marking Tool Mark After Modification
B2	2.5	2.0	2.5
B3	0.5	1.0	1.5
B4	2.0	1.0	2.0
B15	1.0	0	0.5
B19	3.0	2.0	2.5
B26	4.5	3.5	4.0

Table 7.4: Individual results for Substitute Marking Process

	Corpus B Unmodified Marking Tool Result	Corpus B Modified Marking Tool Result	Corpus B Human Mark
Mean	2.1333	2.2500	2.4667
Standard Deviation	1.1059	1.0647	1.1290

Table 7.5: Comparison of the mean and standard deviation of the marking tool results and the human mark

	Corpus B Unmodified Tool Result	Corpus B Modified Tool Result
Percentage Marked Correctly (number marked correctly / number marked)	30% (9/30)	37% (11/30)
Pearson Correlation Coefficient r	0.9013	0.9108
Mean of Absolute Value of Deltas	0.467	0.383

Table 7.6: Percentage marked correctly and Pearson Correlation Coefficient for Corpus B

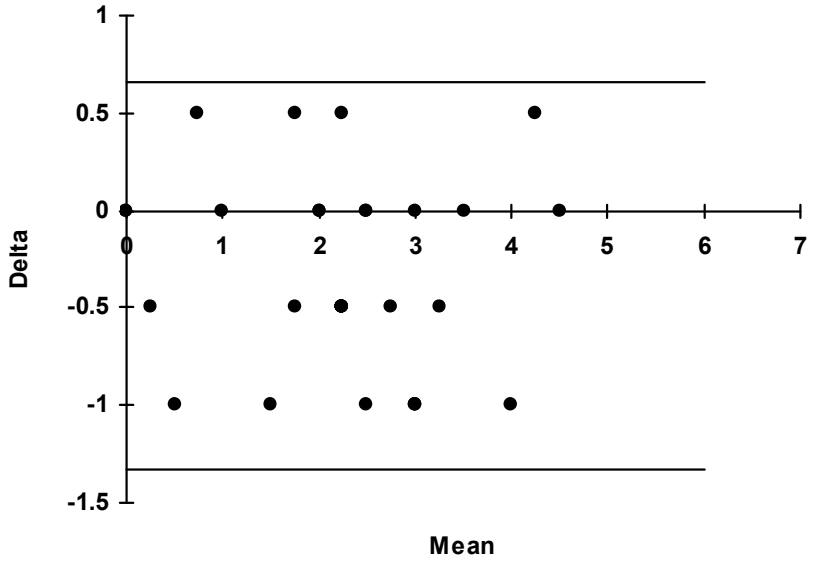

Figure 7.1: Bland-Altman plot of Corpus B data using unmodified marking tool

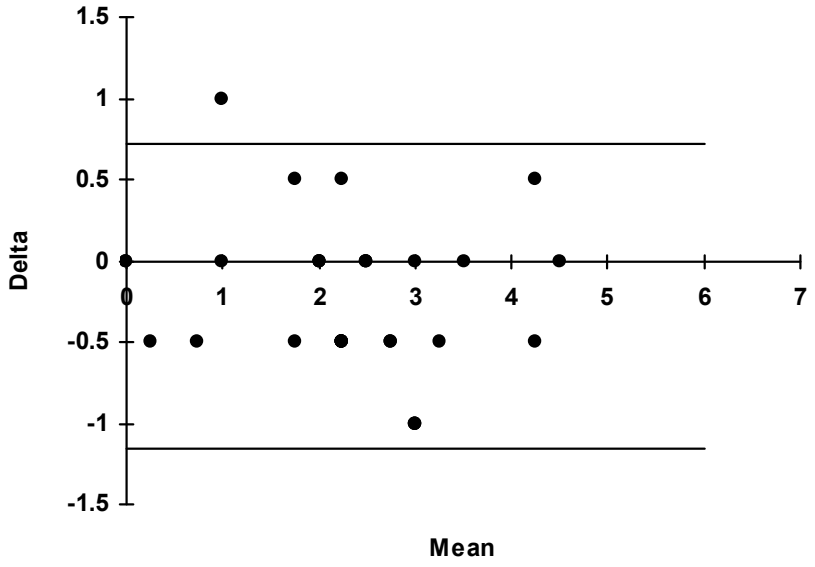

Figure 7.2: Bland-Altman plot of Corpus B data using modified marking tool

Measure for Score(Best) – Human Mark	Corpus B Unmodified Tool Result	Corpus B Modified Tool Result
Mean	-0.3333	-0.2167
Standard Deviation	0.4971	0.4676
Upper 95% Confidence Level	0.6609	0.7185
Lower 95% Confidence Level	-1.3275	-1.1519

Table 7.7: Descriptive statistics for the Bland-Altman plots for Corpus B data

In all cases, the results from the modified marking tool show an improvement, with the mean being closer to that of the human score, a larger number of diagrams marked correctly, and the Pearson correlation coefficient closer to 1.

Figure 7.1 and Figure 7.2 show the Bland-Altman plots for the unmodified and modified marking tools. Visual inspection shows an improved symmetry about the y = 0 axis, and an improved distribution, in that there are fewer points with y values

of +1 and -1.

Data within Table 7.7 confirms these observations showing that the Bland-Altman mean is closer to zero and the 95% confidence band is narrower for the modified tool.

7.2 Incorrect Relationship Matching for Corpus A Diagrams

As described in Chapter 6.3.2, when marking diagrams, the marking tool sometimes matches relationships between the student and model answer diagrams which a human marker would not match. This always leads to the marking tool awarding a higher mark than the human marker.

The approach taken was to try to design a computational test which would identify diagrams which fall within this category, and then design a process for correcting the mark. The alternative strategy would have been to modify the evaluation and marking processes for all diagrams. But with an initial result of 64% of diagrams marked correctly, this alternative strategy would introduce the risk of degrading the overall result whilst fixing the 6 diagrams identified with this problem.

Three options were investigated to resolve this issue. All of these are described in the following sections.

7.2.1 Use of Marking Tool Entity Matching Results

Within the marking tool an array bestEntityMatch is produced as the marking tools best attempt at matching entities between the student answer and model answer diagrams. If this match was accurate then the data could be used to detect and disqualify relationships connected to the wrong entity.

Entity on Student Diagram 10051 (Corpus A)	Matched Entity on Specimen Solution Diagram 0
Prerequisite Course	Prerequisite
Client	Client
course	No Match
UNIT	Unit
Employee	Employee
Entity on Student Diagram 10051 (Corpus A)	**Matched Entity on Specimen Solution Diagram 1**
Prerequisite Course	Course
Client	Client
course	No Match
UNIT	Unit
Employee	Employee

Table 7.8: Marking tool entity matches for student answer diagram 10051 from Corpus A

Unfortunately, this data was found to be very inaccurate. As can be seen in Table 7.8, the marking tool's inability to correctly identify entity "course" and to differentiate between "Prerequisite" and "Course" made this data unsuitable to use as an effective fix for this problem.

7.2.2 Entity Node Consistency Checking Method

This method was derived from two factors:

- The marking tool is unable to correctly identify entity names.
- Looking at student answer A51 (see Figure 6.2), entity "course" has 4 relationships connected to it, but when translating the relationship matching, there is no entity on the specimen solution that has those 4 same relationships connected.

Given this, a process was derived to treat the entities as unnamed nodes, to create

tables for the student answer and specimen solution diagrams showing the relationships connected to each node, then checking for discrepancies between the student answer and the specimen solution.

The following is a worked example using diagram 91 from Corpus A (see Figure 7.3). Firstly, the student answer diagrams are parsed so as to produce groupings of the relationships intersecting at each node. See Table 7.9: Node mapping for diagram A91.

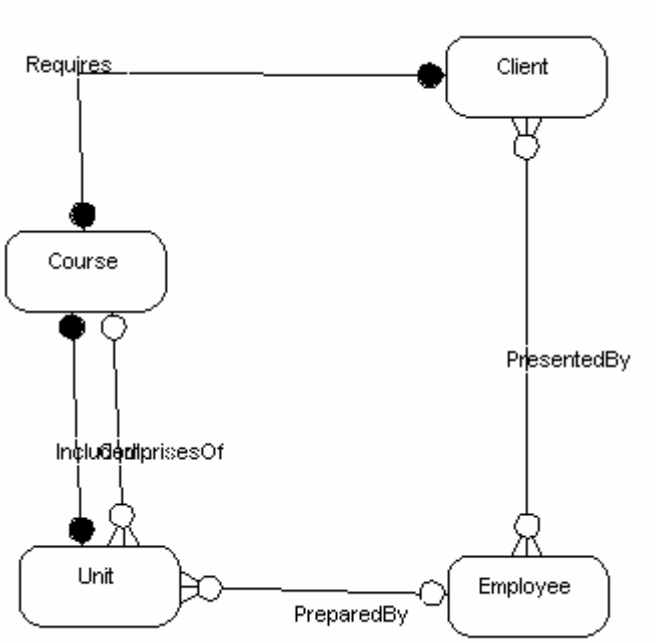

Figure 7.3: Student answer diagram 91 from Corpus A

Node	Relationships
Course	IncludedIn, ComprisesOf, Requires
Unit	IncludedIn, ComprisesOf, PreparedBy
Employee	PreparedBy, PresentedBy
Client	Requires, PresentedBy

Table 7.9: Node mapping for diagram A91

Specimen Solution Relationship	A91 Relationship Matched by Marking Tool
Receives	-
Gives	PresentedBy
Presents	Requires
Has	-
Is	ComprisesOf
Prepares	PreparedBy
BelongsTo	IncludedIn

Table 7.10: Marking tool relationship mapping for A91

Specimen Solution Node	Relationships
Prerequisite	ComprisesOf
Course	ComprisesOf, Requires, IncludedIn
Unit	IncludedIn, PreparedBy
Employee	PreparedBy, PresentedBy
Presentation	Requires, PresentedBy
Client	-

Table 7.11: Node mapping for specimen solution using diagram A91 substitutions

Next, the process is repeated for the specimen solution, but with the relationship in the specimen solution substituted by the relationship that has been matched in the student answer.

Then, a node map for the specimen solution is prepared, substituting the A91 student answer relationships with the specimen solution relationships (Table 7.10). Table 7.11: Node mapping for specimen solution using diagram A91 substitutions, is produced.

Finally, a comparison of the two node mapping tables is done ignoring entity names. So each row of Table 7.9 is compared with Table 7.11 until a matching node containing at least the same relationships is found. If none is found then the test has failed and the diagram's mark is suspect. As an example, the A91 node at entity

"Employee" contains relationships "PreparedBy, PresentedBy". This node passes the test because the specimen solution node at entity "Employee" contains at least the same relationships. However, diagram A91 fails this test as a whole because the node at entity "Unit" contains relationships "IncludedIn, ComprisesOf, PreparedBy". There is no node in Table 7.11 with these relationships.

This test was implemented within a class called NodeCheck and the Java code for this class is included within the appendices.

Testing of the 6 diagrams under investigation showed that 3 passed this test and 3 failed. The test is only effective on those diagrams that fail (see Table 7.12).

Diagram	Pass / Fail
A51	Fail
A84	Pass
A91	Fail
A94	Fail
A135	Pass
A152	Pass

Table 7.12: Results of the entity node consistency test

The second part of the investigation was to define a corrective action for those diagrams which failed the previously defined test. Two options were considered.

The first option considered was to use a similar method as above, but to disqualify relationships until the test was passed. A mark could then be obtained with the disqualified relationships removed. Whilst this method would always reduce the mark, analysis showed that sometimes the wrong relationship would be removed. Also, in some cases, several permutations of disqualified relationships could pass the

test and there was no sound method to define the correct relationship to be removed.

The second option considered was determined from an experiment on the 6 diagrams which showed that changing the parameter values and retesting these 6 diagrams significantly improved the marking tool results. To make use of this phenomenon, code was implemented to dynamically change the parameter values. On detection of a failure from the test, the parameter values are changed to the new ones and the diagram is marked again to obtain the score. Parameter values are returned to their original values after marking. Whilst working empirically, there was no evidence which suggested that this method was a sound way to obtain the correct mark.

Considering the low detection success rate (50%) and the lack integrity of the methods to achieve a modified mark, this solution was not adopted to fix the incorrect relationship matching.

7.2.3 Entity Exact Name Matching

Given the problems with the Marking Tool Entity Matching and the Entity Node Consistency Checking methods, these two methods were disregarded and a completely new method was formulated. This method was derived from the observation that for the 6 diagrams exhibiting this problem, most of the entities have exactly the same name as the entities on the specimen solution. This led me to the hypothesis that the natural thought process of the human marker might be to put a heavy emphasis on matching entities with identical names. Taking diagram A51 as an example, the student has inserted a relationship “requires” between entities “client” and “course”. But the specimen solution has no relationships between these 2 entities.

If my hypothesis is correct, then for the human marker there is no doubt that “client” and “course” on the student answer are exactly the same as “Client” and “Course” on the specimen solution because they have exactly the same name. Hence the human marker would award no marks for the student relationship “requires”.

This hypothesis led to the implementation of a new process as follows.

Firstly, a mapping is produced which maps all entities on the specimen solution to those entities with exactly the same name (ignoring case) in the student answer.

Then, taking each relationship in turn, 2 tests are performed.

1. For each entity at either end of the relationship on the student answer, if either (or both) of those entities are contained within the mapping, then the equivalent relationship within the specimen solution must connect to the mapped entity in the specimen solution.

2. Going to the equivalent relationship on the specimen solution first and taking each entity at the end of this relationship, if either (or both) of those entities are contained within the mapping, then the relationship on the student answer must connect to the mapped entities.

If any relationship fails either of these tests then the relationship is disqualified from scoring.

Two examples are needed to demonstrate these rules. Firstly, on diagram A51 (Figure 6.2), relationship “requires” is equated with “Receives” on the specimen solution. Entities “client” and “course” are directly mapped to “Client” and “Course” on the specimen solutions. To pass test 1 the relationship “Receives” should be

connected between “Client” and “Course” on the specimen solutions. It doesn’t so this test fails. Applying test 2, relationship “Receives” connects “Client” and “Presentation”. Entity “Presentation” doesn’t appear in the mapping (no direct equivalent on the student answer) but “Client” does. So to pass test 2 the relationship “requires” should be connected to “client”. It does and this relationship passes test 2.

On diagram A84 (Figure 7.4), the relationship “presentsOf” is equated to “Presents” in the specimen solutions. Using the same logic as above, entity “Presenters” doesn’t appear in the mapping and so this relationship passes test 1. However, both specimen solution entities “Presentation” and “Course” do appear in the mapping and this relationship fails test 2.

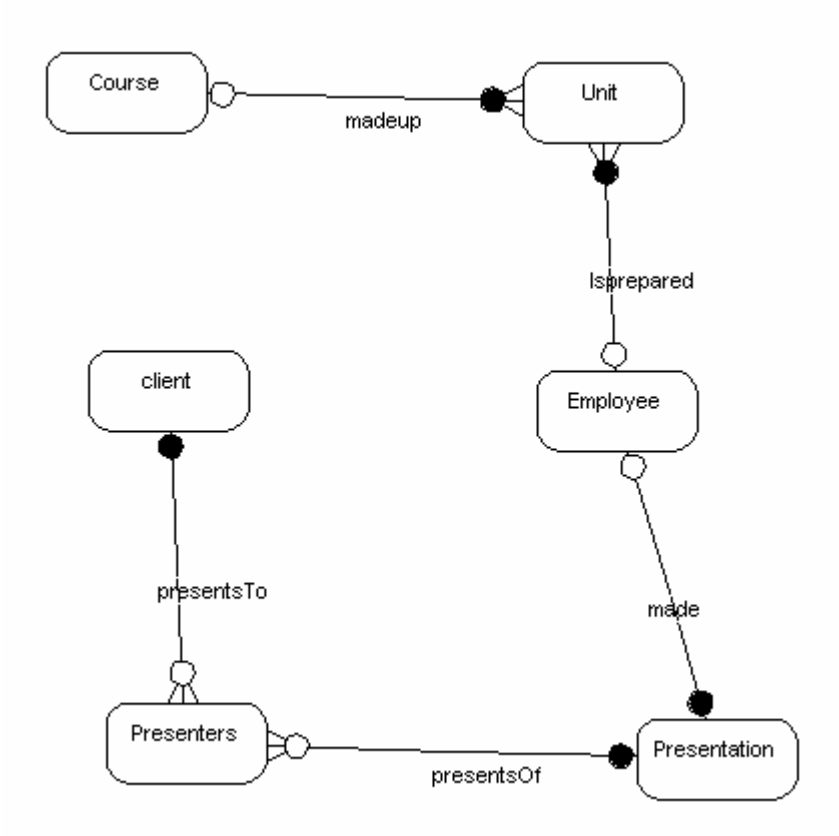

Figure 7.4: Student answer diagram 10084 from Corpus A

This process was implemented within a class called EntityNameCompare and the Java code for this class is included within the appendices.

Table 7.13 shows the results of all diagrams which had their mark changed after implementation of this modification. As can be seen, 10 diagrams were affected by this modification, and of these, 8 had their marks improved and 2 had theirs degraded (further away from the human mark). All 6 of the diagrams identified in root cause analysis were improved and 4 of these were corrected to equal the human mark.

Diagram	Human Mark	Marking Tool Mark Prior to Modification	Marking Tool Mark After Modification
A14	1.5	2.5	2.0
A51	1.0	2.0	1.0
A84	2.0	3.0	2.0
A91	0.5	1.5	1.0
A94	5.0	6.0	5.0
A107	3.5	3.5	3.0
A135	0.5	1.5	1.0
A142	2.0	1.5	1.0
A152	2.0	3.0	2.0
A183	4.0	4.5	4.0

Table 7.13: Individual results for Entity Exact Name Matching Process

Given these results, this process and implementation was adopted as the final solution for this problem.

7.2.4 Test Results

For the reasons given in Section 7.3, all of the following data is taken using the parameter Synonym Check set to false.

	Corpus A Unmodified Marking Tool Result	Corpus A Modified Marking Tool Result	Corpus A Human Mark
Mean	3.236	3.201	3.142
Standard Deviation	1.496	1.509	1.549

Table 7.14: Comparison of the mean and standard deviation of the marking tool results and the human mark

	Corpus A Unmodified Tool Result	Corpus A Modified Tool Result
Percentage Marked Correctly (number marked correctly / number marked)	65% (128/197)	67% (132/197)
Pearson Correlation Coefficient r	0.968	0.973
Mean of Absolute Value of Deltas	0.221	0.195

Table 7.15: Percentage marked correctly and Pearson Correlation Coefficient for Corpus A

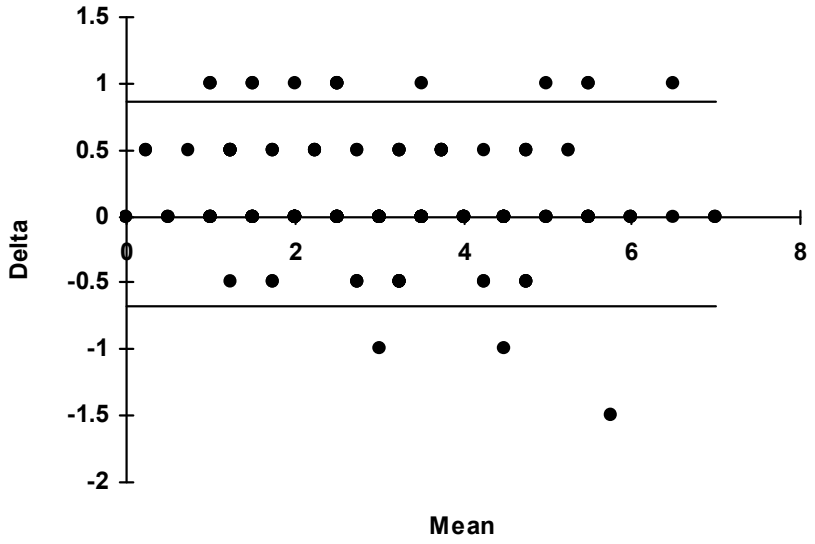

Figure 7.5: Bland-Altman plot of Corpus A data using unmodified marking tool

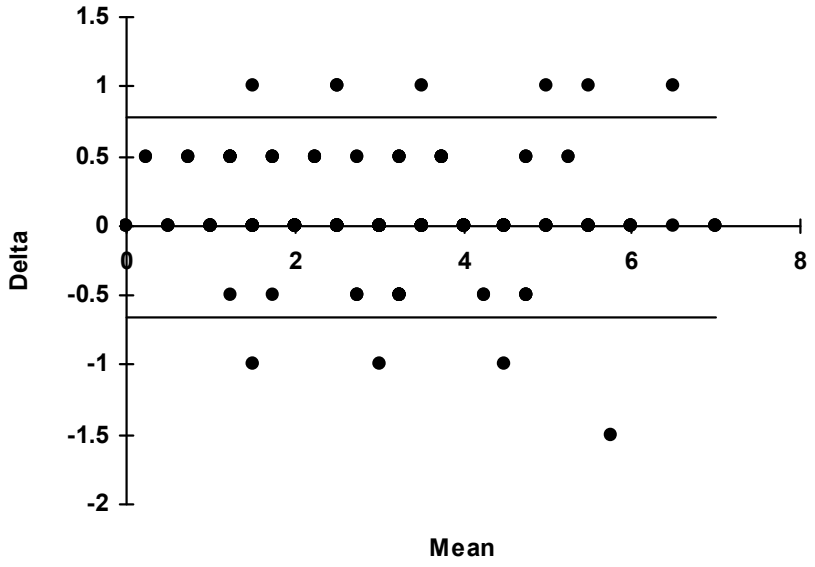

Figure 7.6: Bland-Altman plot of Corpus A data using modified marking tool

Measure for Score(Best) – Human Mark	Corpus A Unmodified Tool Result	Corpus A Modified Tool Result
Mean	0.0939	0.0584
Standard Deviation	0.388	0.3577
Upper 95% Confidence Level	0.8699	0.7738
Lower 95% Confidence Level	-0.6821	-0.657

Table 7.16: Descriptive statistics for the Bland-Altman plots for Corpus A data

Table 7.14, Table 7.15, Figure 7.5, Figure 7.6, and Table 7.16 together give data comparing marking tool performance before and after modification with the Entity Exact Name Matching process. All data appears to demonstrate a measurable improvement in marking tool performance when used for marking Corpus A.

These results are very pleasing in that the effect of the modification is very focused (the modification only affects a small number of diagrams, most of which are

incorrectly marked and the modification improves the mark of a very high proportion of diagrams affected). These results go some way towards validating my original hypothesis as to the reason why this method might work successfully. I believe that the performance improvement achieved justifies further investigations into this aspect of the human marking mechanism and potential further refinement of this modification.

7.3 Tool Relationship and Entity Name Recognition

Root cause analysis identified 4 incorrectly marked diagrams for which improvement was achieved by changing the naming of a single entity or relationship.

Further analysis showed that for 2 of these diagrams, A1 and A24, these are correctly marked when retested with the marking tool Synonym Check parameter set to false. Thomas (2008) advised that the Synonym Check parameter invokes legacy code which was written for a specific situation and recommended that this parameter should be set to false. This being the case, only 2 diagram instances of this phenomenon remained.

Given this situation, I believe that rectification of this issue would only offer a minor improvement to the marking tool performance, and given the activity on other issues and the time constraints of this project, no further investigation of this issue was investigated. I don't feel that this decision has had any impact on meeting the objectives of this project.

7.4 Decomposed m:n Relationship Recognition

For this problem, it was found on diagram 24 from Corpus B that a decomposed m:n

relationship was not correctly recognised by the marking tool. The relationships on the student diagram answer were not named. It was found that by editing the student diagram and giving the two relationships the same name as the decomposition entity on the student answer, that this problem with the marking tool was eliminated.

Once again, due to time constraints of this project, no fix for this problem was implemented. However, my examination of the marking tool code leads me to believe that an implementation of the fix of adding relationship naming on decomposed m:n relationships is feasible.

7.5 Summary

The root cause analysis identified candidate issues to fix and the 2 most significant of these were addressed.

The process used to investigate, design, implement and test fixes was broadly similar for the 2 issues addressed. In both cases the incorrectly marked diagrams which exhibit the issue were reviewed in order to identify the characteristics of the incorrectly marked diagrams and to formulate a theory on the difference between the methods of the human marker and the marking tool.

There are some common features of the results achieved by these 2 marking tool modifications:

- Both modifications achieved a significant improvement to the marking tool performance as demonstrated by the statistical data.

- Both modifications had a very focussed impact on the marking performance. For each modification a very high proportion of the marks affected by the

modifications were incorrect. After tool modification, a very high proportion of affected marks were improved.

Chapter 8 Results for Extended Corpus A

On completion of the marking tool improvements, an additional set of diagrams was made available as an extension to Corpus A. This additional set of diagrams consisted of 195 additional student answer diagrams together with the human marker grades.

Two tests were performed on these diagrams. Firstly, the marking tool was used in its original state (with my improvements disabled) to obtain a set of results. Then the test was repeated with my improvements enabled. For this experiment I only used my Entity Exact Name Matching improvement, as described in section 7.2.3 since my other improvement (described in section 7.1) was designed for Corpus B type diagrams only.

This test was expected to indicate if the marking tool improvements are of general usefulness and significance or if the improvements are specific to a small number of diagrams within the original Corpus A

Note that all of this testing was performed with parameter Synonym Check set to false. Also, the results that follow relate only to the new diagrams (the testing was not performed on Corpus A in its entirety).

Table 8.1 shows the diagrams from the Extended Corpus A set whose mark was affected by the Entity Exact Name Matching modification. Of 19 diagrams whose mark was affected, 3 of the diagrams had their marks degraded (further away from the human mark), and 15 had their marks improved by the marking tool modification.

Diagram	Human Mark	Marking Tool Mark Prior to Modification	Marking Tool Mark After Modification
Ouq15scan10201	4.0	5.0	3.0
Ouq15scan10210	2.5	2.5	2.0
Ouq15scan10226	3.5	3.5	3.0
Ouq15scan10232	1.0	1.5	1.0
Ouq15scan10233	1.0	3.0	2.0
Ouq15scan10240	0.5	1.0	0.5
Ouq15scan10252	2.0	4.0	2.0
Ouq15scan10254	2.0	1.5	1.0
Ouq15scan10267	2.5	3.0	2.5
Ouq15scan10277	3.0	4.5	3.5
Ouq15scan10284	1.0	2.0	1.5
Ouq15scan10297	3.0	4.0	3.0
Ouq15scan10298	2.0	2.5	2.0
Ouq15scan10308	1.0	2.0	1.5
Ouq15scan10315	3.5	5.0	4.0
Ouq15scan10320	3.5	4.0	3.5
Ouq15scan10340	1.5	2.0	1.5
Ouq15scan10348	1.5	2.0	1.5
Ouq15scan10396	4.0	5.0	4.0

Table 8.1: Comparison of individual diagram marks for Extended Corpus A

	Extended Corpus A Unmodified Marking Tool Result	Extended Corpus A Modified Marking Tool Result	Extended Corpus A Human Mark
Mean	3.482	3.405	3.387
Standard Deviation	1.528	1.549	1.584

Table 8.2: Comparison of the mean and standard deviation of the marking tool results and the human mark

	Extended Corpus A Unmodified Tool Result	Extended Corpus A Modified Tool Result
Percentage Marked Correctly (number marked correctly / number marked)	60% (117/195)	64% (125/195)
Pearson Correlation Coefficient r	0.948	0.963
Mean of Absolute Value of Deltas	0.290	0.238

Table 8.3: Percentage marked correctly and Pearson Correlation Coefficient for Extended Corpus A

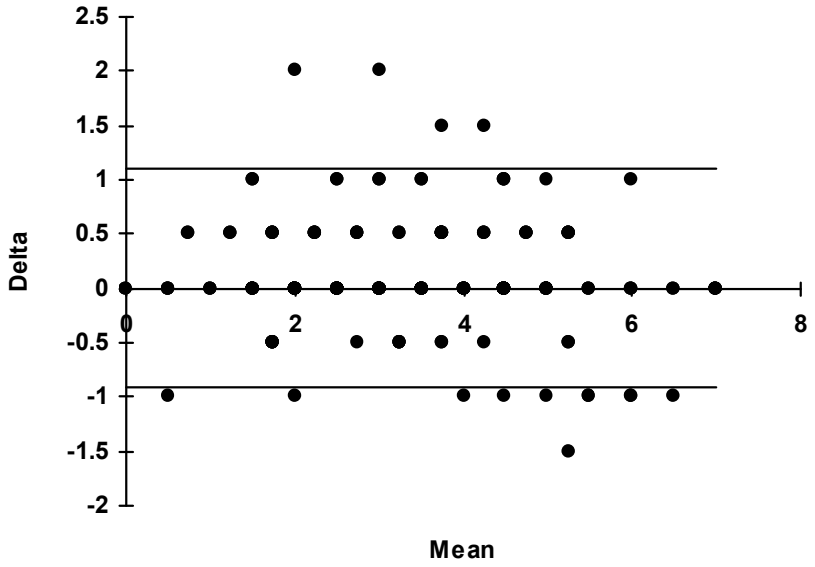

Figure 8.1: Bland-Altman plot for Extended Corpus A using unmodified marking tool

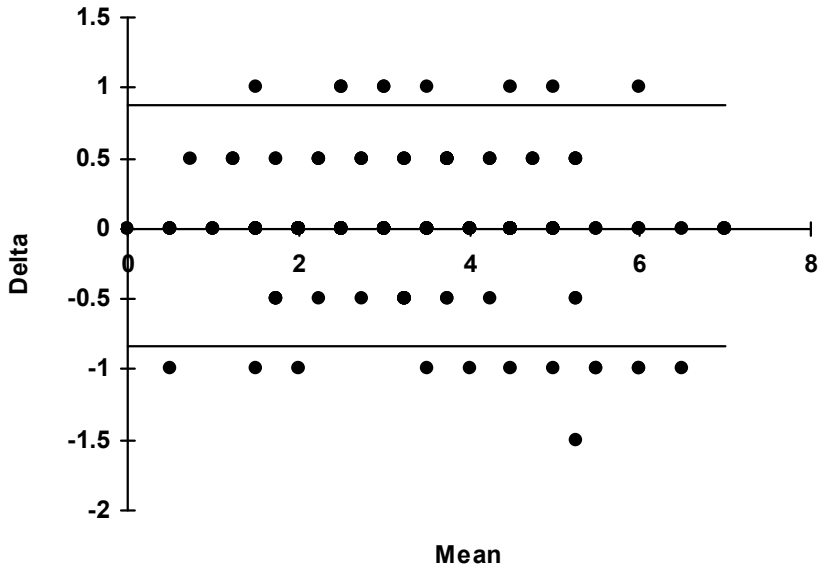

Figure 8.2: Bland-Altman plot for Extended Corpus A using modified marking tool

Measure for Score(Best) – Human Mark	Extended Corpus A Unmodified Tool Result	Extended Corpus A Modified Tool Result
Mean	0.0949	0.0179
Standard Deviation	0.500	0.426
Upper 95% Confidence Level	1.095	0.870
Lower 95% Confidence Level	-0.905	-0.834

Table 8.4: Descriptive statistics for the Bland-Altman plots for Extended Corpus A data

All results in Table 8.2, Table 8.3 and Table 8.4, and the Bland-Altman plots of Figure 8.1 and Figure 8.2 show that my marking tool modification had a significant effect on the marking tool performance. Comparing these data with the marking tool data for the original Corpus A diagram set (in Section 7.2.4):

- The data for marking tool performance of the unmodified marking tool shows that

the unmodified marking tool marks the Extended Corpus A diagram set less accurately than it did for the original Corpus A diagram set. As an example, Corpus A is marked with an accuracy of 65% whereas the Extended Corpus A is marked with an accuracy of 60%. Other data measures indicate the same phenomenon.

- My marking tool modification has an effect on a far greater proportion of diagrams within the Extended Corpus A diagram set than for the original Corpus A diagrams set. For the Extended Corpus A set, 19 diagrams out of 195 were affected (9.7%). For the original Corpus A set, 10 diagrams out of 197 were affected (5.1%).

These observations indicate that the Extended Corpus A diagram set contains a higher occurrence of the incorrect relationship matching issue than that of the original Corpus A diagram set.

These results are extremely pleasing and reinforce my conclusions that my Entity Exact Name Matching modification is of general usefulness and significance. These results confirm that this method warrants further investigations, both to conclude on my hypothesis on the underlying marking mechanism and to investigate if any further refinement is possible.

Chapter 9 Conclusions

As a conclusion to this project I have reviewed the conduct and effectiveness of the project in the light of the aims and objectives that were set at the start.

9.1 Project Review

In reviewing the project I have categorised the project aims and objectives into 3 broad areas.

9.1.1 Evaluation and Analysis of Marking Tool Performance

The first of the aims and objectives was to produce a controlled experimental method so as to ensure that results were reproducible. This was achieved by automating the setting of parameters and by automating the data logging.

Experiments were performed to investigate the optimal values of the user defined parameters. These were adequate for the purposes of this project but were not fully comprehensive. A full investigation into a process for optimising these parameters, and whether the optimised parameter values differ for different E-R diagram questions could potentially be the subject of another research project.

The second of the aims and objectives was to define statistical analysis methods for measuring the marking tool performance. Previously, researchers had relied heavily on the Pearson correlation coefficient together with scatter plots. The Bland-Altman plots were found to significantly improve on the scatter plots by increasing sensitivity of the y-axis and by automatically providing a value of disagreement. Descriptive statistics which accompany the Bland-Altman plots give useful data for analysing

trends. Within this dissertation, I adopted the Bland-Altman plot method to report my results in preference to scatter plots. This study completed the third objective of this project.

However, limitations to the statistical methods remain as follows:

- Both Bland-Altman and scatter plots are incomplete graphical representations as both suffer from having multiple data at the same graphical point.
- No one statistical measure was found to provide a complete picture of marking tool performance, hence all results were presented using a number of different measures.
- Measures were found to be quite insensitive to small improvements in marking tool performance. The most sensitive measure was found to be the mean of the absolute value of deltas.
- When performing a marking tool modification, it was found that there was a need to report the "before" and "after" results for the diagram marks affected by the modification.

Thomas (2008) advised me of the AC_1 statistic (Gwet, 2001) which he believes to be more appropriate than the Pearson Correlation Coefficient for this task. This statistic may prove useful for future projects.

9.1.2 Root Cause Analysis

Two major impediments were encountered during the root cause analysis which led to a large number of diagrams being categorised with Human Error or Indeterminate

causes. This resulted in a significant data loss for this project as approximately 70% of diagrams analysed fell into these 2 categories.

The first of these impediments was lack of detail behind the human mark. For a number of diagrams it was impossible to determine how the human mark had been determined as there was no traceability to the marks awarded for each relationship. Hence, for a number of diagrams I could not determine if there was human error in the supplied human mark or if there was error in my mark.

A way of resolving this problem for future projects would be to have the human marker fill in a matrix for each marked diagram showing the detail behind the mark awarded.

The second impediment was the lack of a complete set of marking rules. As an example, diagrams A118 and B24 have no names applied to any relationships. If a specific rule were established and documented for marking of unnamed relationships then it would be easy to validate the human mark using that rule, and it would be easy to implement that rule within the marking tool.

The method adopted for root cause analysis was to decompose the marks into the award for each component and to compare human and marking tool marks for each. Despite the limitations described above, this method successfully produced adequate evidence to support the process of designing and implementing marking tool modifications. Given this outcome, my fourth objective was achieved.

9.1.3 Design and Implementation of Modifications

Within Chapter 7, I reported on two major marking tool modifications.

The first of these was the method to address incorrect relationship matching for Corpus B diagrams. From the root cause analysis data, it was observed that in general, where a student had missed a subtype, the human marker would only penalise the student for the missing subtype and would award marks for any correct relationships that would otherwise be connected to the missing subtype. The marking tool was modified to copy this marking practice. Results showed that a significant improvement in marking performance was achieved for Corpus B.

The second major marking tool modification is the Entity Exact Name Matching method which was designed to overcome the incorrect relationship matching issue for Corpus A diagrams. This modification doesn't yet have a sound theoretical basis. My hypothesis is that the human marker places a high emphasis on the entity names where the names in the student answer exactly match those in the specimen solution, and the human marker disqualifies relationships which are not connected to the correct entities. This modification has produced a significant improving of the marking performance for Corpus A diagrams as reported in Section 7.2.4 and Chapter 8. For this modification, there is scope for further research.

- If my theory is correct, then further improvement in the marking tool performance might be achievable by changing the implementation to match nearly identical entity names (e.g. "Course" and "Courses") whereas the current implementation only identifies identical matches.

- The hypothesis of why this modification works so well might be tested by examining all the diagrams which have their marks improved by this modification. In particular, the marking of the Extended Corpus A set has provided a significant new data set to work with.

Together, these modifications represent successful completion of my fifth and final project objective.

9.2 Future research

This project, combined with the earlier work of Thomas et al. (2006b) has demonstrated that the marking tool performance can be improved by iteratively following the process steps of acquiring a corpus of marked diagrams, using the marking tool to mark the corpus, performing root cause analysis where discrepancies exist between the human and marking tool mark, and modifying the software to correct for the discrepancies.

However, the results of this project indicate that simply repeating this process may not result in a marking tool which can universally be used for marking E-R diagrams. Evidence for this is that:

- When used to mark questions requiring different E-R diagram solutions (Corpus A and Corpus B), the marking tool performance was significantly different.

- The results of the root cause analysis for marking discrepancies yielded different failure mechanisms for Corpus A and Corpus B.

- The optimised values of the user set parameters were slightly different between

Corpus A and Corpus B.

These results suggest that each new E-R diagram question may reveal new deficiencies within the marking tool. The worst case scenario is that the marking tool performance needs to be validated (and perhaps modified or optimised) for use with each new E-R diagram question. On the other hand, it could be that the differences in the marking tool response to Corpus A and Corpus B are entirely due to the subtypes in Corpus B and due to the large difference in the sizes of the 2 corpora. The best scenario being that Corpus A and Corpus B together establish the entire range of marking tool deficiencies and fixing these will establish a universal marking tool. The true situation can only be established by further research to investigate the marking tool performance variability over a wider range of E-R diagram questions, into the causes of the marking tool performance variability for different question types, and into alternative methods for measuring and optimising marking tool performance.

References

Attali, Y. and Burstein, J. (2005) *Automated Essay Scoring With E-rater® v.2.0*, Report No. RR-04-45, Educational Testing Service, Princeton, NJ, USA.

Batmaz, F. and Hinde, C.J. (2006) 'A Diagram Drawing Tool for Semi-Automatic Assessment of Conceptual Database Diagrams', *Proceedings of the 10th Annual International Conference in Computer Assisted Assessment*, Loughborough University, Loughborough, UK, pp. 68-81.

Bloom, B. S. (1956) *Taxonomy of Educational Objectives, the classification of educational goals – Handbook I: Cognitive Domain* McKay, New York

Blostein, D. (1996) 'General diagram-recognition methodologies', *Proceedings of Workshop on Graphics Recognition*, 10-11 Aug. 1995, University Park, PA, USA, pp 106-22.

Bull, J. (2001) *TLTP85 Implementation and Evaluation of Computer-assisted Assessment: Final Report.* [Online], http://www.caacentre.ac.uk/dldocs/final_report.pdf [Accessed 26 May 2007].

Carr, R. (2004) *XLStatistics* [Online], http://www.deakin.edu.au/~rodneyc/xlstats.htm. [Accessed 23 April 2007].

Carter, J. and English, J. and Ala-Mutka, K. and Dick, M. and Fone, W. and Fuller, U. and Sheard, J. (2003) 'How Shall We Assess This?' *ACM SIGCSE Bulletin*, Working group reports from ITiCSE on Innovation and technology in computer science education ITiCSE-WGR '03, Volume 35, Issue 4, pp 107-23.

Chok, S. S. and Marriott, K. (1995) "Parsing visual languages". *Proceedings of the Eighteenth Australian Computer Science Conference*, Australian Computer Science Communications, 17, 90-98.

Dallal, G (n.d.) *The Little Handbook of Statistical Practice* [Online], http://www.tufts.edu/%7Egdallal/LHSP.HTM [Accessed 1 March 2007].

Gwet, K. (2001) *Handbook of Inter-Rater Reliability*, STATAXIS publishing Company, USA. ISBN 0-9708062-2-1.

Higgins, C.A. and Gray, G. and Symeonidis, P. and Tsintsifas, A. (2005) 'Automated Assessment and Experiences of Teaching Programming', *Journal on Educational Resources in Computing (JERIC)*, Volume 5, Issue 3 (September 2005), Article No. 5.

Higgins, C. and Hegazy, T. and Symeonidis, P. and Tsintsifas, A. (2001) 'The CourseMarker CBA System: Improvements over Ceilidh', *5th Annual Computer Assisted Assessment Conference*, Loughborough, UK, 2 – 4 July, pp. 189 - 201.

Higgins, C. and Symeonidis, P. and Tsintsifas, A. (2002) 'Diagram-based CBA using DATsys and CourseMaster', *Proceedings of International Conference on Computers in Education, 2002*. 3-6 Dec., Vol 1, pp. 167 – 172.

Hoggarth, G. and Lockyer, M. (1998) 'An Automated Student Diagram Assessment System', *Proceedings of the 6th annual conference on the teaching of computing and the 3rd annual conference on Integrating technology into computer science education*, Dublin City Univ., Ireland, pp. 122 – 124.

Iizuka, K. and Tanaka, J. and Shizuki, B. (2001) 'Describing a Drawing Editor by Using Constraint Multiset Grammars', *Proceedings of the International Symposium on Future Software Technology (ISFST2001)*, Zheng Zhou, China, Nov. 5-8, pp. 119-124.

Jurafsky, D. and Martin, J. H. (2000) *Speech and Language Processing*, New Jersey, Prentice-Hall. ISBN 0-13-122789-X.

Kirkwood, B. and Sterne, J. (2003) *Essential Medical Statistics*, Malden, Blackwell Science.

Manning, C.D. and Schutze, H. (1999) *Foundations of Statistical Natural language Processing*, MIT Press, Cambridge, Massachusetts, USA. ISBN 0-262-13360-1.

Marriott, K. (1994) 'Constraint Multiset Grammars', *Proceedings of the IEEE Symposium on Visual Languages*, St Louis, USA, Oct. 4-7, pp.118 – 125.

Marriott, K and Meyer, B and Wittenburg, B (1998) 'A Survey of Visual Language Specification and Recognition' in Marriott, K. and Meyer, B. (Eds), *Visual Language Theory*, Springer-Verlag, New York, ISBN 0-387-98367-8.

The Open University (1995) M865 *Project Management: Unit 1, Project Initiation*, Milton Keynes, Open University.

Rekers, J. and Schurr, A. (1997) 'Defining and Parsing Visual Languages with Layered Graph Grammars', *Journal of Visual Languages & Computing*, Volume 8, Number 1, February 1997, pp. 27-55(29).

Smith, N. and Thomas, P. and Waugh, K. (2004) 'Interpreting Imprecise Diagrams', Proceedings of the Third International Conference in the Theory and Application of Diagrams. March 22-24, Cambridge, UK. *Springer Lecture Notes in Computer Science*, Eds: Alan Blackwell, Kim Marriott, Atsushi Shimojima, 2980, 239--241. ISBN 3-540-21268-X.

StatSoft Inc. (n.d.) *Basic Statistics* [Online], http://www.statsoft.com/textbook/stbasic.html#Correlationsb [Accessed 9 August 2007].

Tanaka, J. (2001) 'Visual Parsers based on Extended Constraint Multiset Grammars', *Japan-Tunisia Workshop on Informatics (JTWIN 2001)*, Tsukuba, October 25-26, pp. 69-72.

Thomas, P. (2003) *The Evaluation of Electronic Marking of Examinations*, 8th Annual International Conference on Innovation and Technology in Computer Science Education, July 2003, Thesaloniki, Greece

Thomas, P. (2004a) *Grading Diagrams Automatically*, Technical Report No 2004/01, Department of Computing, The Open University, Milton Keynes, UK.

Thomas, P. (2004b) *Comparing machine graded diagrams with human markers: some observations*, Technical Report No 2004/27, Department of Computing, The Open University, Milton Keynes, UK.

Thomas, P. (2004c) *Drawing Diagrams in an Online Examination*, Technical Report No 2004/14, Department of Computing, The Open University, Milton Keynes, UK.

Thomas, P. (2008) Personal communication to the author, 6 February.

Thomas, P. and Smith, N. and Waugh, K. (2006b) *An approach to the automatic grading of imprecise diagrams*, Technical Report No 2006/16, Department of Computing, The Open University, Milton Keynes, UK.

Thomas, P. and Waugh, K. and Smith, N. (2005) 'Experiments in the Automatic Marking of ER-Diagrams', *ACM SIGCSE Bulletin, Proceedings of the 10th annual SIGCSE conference on Innovation and technology in computer science education ITiCSE '05*, Volume 37 Issue 3.

Thomas, P. and Waugh, K. and Smith, N. (2006a) 'Using Patterns in the Automatic Marking of ER-Diagrams', *ACM SIGCSE Bulletin , Proceedings of the 11th annual SIGCSE conference on Innovation and technology in computer science education ITICSE '06*, Volume 38 Issue 3.

Thomas, P.G., Waugh, K., and Smith, N. (2007) 'Tools Learning and automatically assessing graph-based diagrams', *Research Proceedings of ALT-C 2007*, Nottingham, 4-6 September, 2007, pp. 61-74.

Tsintsifas, A. (2002) *A Framework for the Computer Based Assessment of Diagram Based Coursework* Ph.D. Thesis, School of Computer Science and Information Technology, University of Nottingham, UK.

Warburton, W. I. and Conole, G. C. (2003) 'Key Findings from recent literature on Computer-aided Assessment.' In*, ALT-C 2003,* Sheffield*,* 8-10 Sep 2003. Sheffield, UK. [Found online], http://eprints.soton.ac.uk/14113/ [Accessed 26 May 2007].

Waugh, K. and Thomas, P. and Smith, N. (2004) 'Toward the Automated Assessment of Entity-Relationship Diagrams', *Learning and Teaching Support Network – Information and Computer Science (LTSN-ICS), Teaching, Learning and Assessment of Databases (TLAD) second workshop*, Heriot-Watt, Edinburgh, July 2004.

Index

Appendix A: Code Fixes to Correct Bugs within the Marking Tool

The following code change was implemented to ensure that the best score is returned in cases where 2 solution diagrams are being marked.

Within MarkingToolFrame class, the following code …

```
// Find best auto score
double best = marks[ans][noOfSolutions+1];
for (int sol=1; sol<noOfSolutions; sol++) {
  if (marks[ans][sol+1] > best) {
    best = marks[ans][sol+1];
  }
}
```

was replaced with the following …

```
// Find best auto score
    double best = marks[ans][noOfSolutions+1];
    for (int sol=1; sol<noOfSolutions; sol++) {
      if (marks[ans][sol+noOfSolutions+1] > best) {   // sol+1 replaced by
sol+noOfSolutions+1
        best = marks[ans][sol+noOfSolutions+1];      // sol+1 replaced by
sol+noOfSolutions+1
      }
    }
```

The following code change was implemented to ensure that the "Relation Name Weight" parameter can be adjusted from the marking tool dialogue box.

Row 124 in the ParameterDialog class was incorrect. The fragment …

```
case 5:
  value = Double.parseDouble(data);
  if (0 <= value && value <= 1.0){
    Parameter.relEntityNameWeight = value;
    }
```

Was corrected to …

```
case 5:
  value = Double.parseDouble(data);
  if (0 <= value && value <= 1.0){
    Parameter.relNameWeight = value;
  }
```

Appendix B: Results for Corpus A Prior to Tool Modification

Student Diagram	Human Mark	Score(0)	Score(1)	Score(Best)
ouq15scan10001jpg	6.5	4	3	4
ouq15scan10002jpg	3.5	3	3	3
ouq15scan10003jpg	6.5	5	4	5
ouq15scan10004jpg	4	3.5	4	4
ouq15scan10005jpg	3	3	3	3
ouq15scan10006jpg	2	2	2	2
ouq15scan10007jpg	2	2.5	2.5	2.5
ouq15scan10008jpg	6	6	5	6
ouq15scan10009jpg	4	4	4	4
ouq15scan10010jpg	4.5	4.5	4	4.5
ouq15scan10011jpg	4.5	4.5	5	5
ouq15scan10012jpg	3.5	3.5	3	3.5
ouq15scan10013jpg	3.5	4	3.5	4
ouq15scan10014jpg	1.5	2	2.5	2.5
ouq15scan10015	1.5	2	2	2
ouq15scan10016jpg	2	2	2	2
ouq15scan10017jpg	4.5	4	4	4
ouq15scan10018jpg	5	5	5	5
ouq15scan10019jpg	2	2	2	2
ouq15scan10020jpg	3.5	4	3	4
ouq15scan10021jpg	4.5	4.5	4.5	4.5
ouq15scan10022jpg	3.5	3	2	3
ouq15scan10023jpg	0	0.5	0.5	0.5

ouq15scan10024jpg	3.5	2.5	2.5	2.5
ouq15scan10025jpg	2.5	2.5	2.5	2.5
ouq15scan10026jpg	2.5	2.5	2.5	2.5
ouq15scan10027jpg	3	3	3	3
ouq15scan10028jpg	2	2	1.5	2
ouq15scan10029jpg	4	4	3	4
ouq15scan10030jpg	2	2	1.5	2
ouq15scan10031jpg	3	3	3	3
ouq15scan10032jpg	1.5	1.5	1.5	1.5
ouq15scan10033jpg	2	2	2	2
ouq15scan10034jpg	4.5	4.5	5.5	5.5
ouq15scan10035jpg	3.5	3.5	3	3.5
ouq15scan10036jpg	1.5	1.5	2	2
ouq15scan10037jpg	3	2.5	1.5	2.5
ouq15scan10038jpg	3.5	3.5	3.5	3.5
ouq15scan10039jpg	4.5	4.5	4.5	4.5
ouq15scan10040jpg	3.5	3.5	3.5	3.5
ouq15scan10041jpg	3	2.5	3	3
ouq15scan10042jpg	1.5	1.5	1.5	1.5
ouq15scan10043jpg	2	2	2	2
ouq15scan10044jpg	1	1	1	1
ouq15scan10045jpg	1	1.5	1.5	1.5
ouq15scan10046jpg	4	4	4	4
ouq15scan10047jpg	1	1.5	1.5	1.5
ouq15scan10048jpg	0	0	0	0
ouq15scan10049jpg	1	1	1.5	1.5

ouq15scan10050jpg	3	3	3.5	3.5
ouq15scan10051jpg	1	1	2	2
ouq15scan10052jpg	2	2	2	2
ouq15scan10053jpg	2	2.5	2.5	2.5
ouq15scan10054jpg	1	1	1	1
ouq15scan10055jpg	3.5	3.5	3.5	3.5
ouq15scan10056jpg	2.5	2.5	2.5	2.5
ouq15scan10057jpg	3	3	3	3
ouq15scan10058jpg	3	3	3	3
ouq15scan10059jpg	3.5	3.5	3.5	3.5
ouq15scan10060jpg	4.5	4.5	5	5
ouq15scan10061jpg	2.5	2.5	2.5	2.5
ouq15scan10062jpg	2	2	2	2
ouq15scan10063jpg	3.5	3.5	3	3.5
ouq15scan10064jpg	5	5	4	5
ouq15scan10065jpg	1.5	1.5	1.5	1.5
ouq15scan10066jpg	7	7	5	7
ouq15scan10067jpg	2.5	2.5	2.5	2.5
ouq15scan10068jpg	3.5	3.5	3.5	3.5
ouq15scan10069jpg	5	5	4.5	5
ouq15scan10070jpg	2.5	2.5	2.5	2.5
ouq15scan10071jpg	3	3	3	3
ouq15scan10072jpg	5.5	5.5	5	5.5
ouq15scan10073jpg	4	4	4	4
ouq15scan10074jpg	3.5	3.5	3	3.5
ouq15scan10075	3.5	1.5	2.5	2.5

ouq15scan10076jpg	3	3	2.5	3
ouq15scan10077jpg	4.5	4.5	4.5	4.5
ouq15scan10078jpg	3	3	3	3
ouq15scan10079jpg	1	1.5	1.5	1.5
ouq15scan10080jpg	2	2	2	2
ouq15scan10081jpg	2	1.5	1.5	1.5
ouq15scan10082jpg	3	3	3	3
ouq15scan10083jpg	3	3	3	3
ouq15scan10084jpg	2	3	3	3
ouq15scan10085jpg	4.5	4.5	4.5	4.5
ouq15scan10086jpg	3.5	3.5	4	4
ouq15scan10087jpg	3.5	3.5	3.5	3.5
ouq15scan10088jpg	4.5	4.5	3.5	4.5
ouq15scan10089jpg	5	4.5	4.5	4.5
ouq15scan10090jpg	1.5	1.5	1.5	1.5
ouq15scan10091jpg	0.5	1.5	1	1.5
ouq15scan10092jpg	3	2.5	2.5	2.5
ouq15scan10093jpg	2.5	2.5	3	3
ouq15scan10094jpg	5	6	5	6
ouq15scan10095jpg	3.5	3.5	3.5	3.5
ouq15scan10096jpg	1.5	1.5	1.5	1.5
ouq15scan10097jpg	6	6	4	6
ouq15scan10098jpg	3	3	3	3
ouq15scan10099jpg	3.5	3.5	4	4
ouq15scan10100jpg	5	5.5	4.5	5.5
ouq15scan10101jpg	5.5	5.5	5	5.5

ouq15scan10102jpg	2	2	2	2
ouq15scan10103jpg	5.5	5.5	5	5.5
ouq15scan10104jpg	0.5	0.5	0.5	0.5
ouq15scan10105jpg	4	4	4	4
ouq15scan10106jpg	3	2.5	3	3
ouq15scan10107jpg	3.5	3.5	3	3.5
ouq15scan10108jpg	2	2	3	3
ouq15scan10109jpg	2	2	2	2
ouq15scan10110	6	5	7	7
ouq15scan10111jpg	4.5	4.5	3.5	4.5
ouq15scan10112jpg	2.5	2.5	2.5	2.5
ouq15scan10113jpg	5.5	5.5	4.5	5.5
ouq15scan10114jpg	4	4	4	4
ouq15scan10115jpg	3	2.5	2	2.5
ouq15scan10116jpg	2.5	2	2	2
ouq15scan10117jpg	4	4	3.5	4
ouq15scan10118jpg	1	2	1.5	2
ouq15scan10119jpg	1	1	1	1
ouq15scan10120jpg	7	7	5	7
ouq15scan10121jpg	1	1	1.5	1.5
ouq15scan10122jpg	5	5	4	5
ouq15scan10123jpg	2	2.5	2.5	2.5
ouq15scan10124jpg	2	2	2	2
ouq15scan10125jpg	4	4	4	4
ouq15scan10126jpg	2.5	2.5	2	2.5
ouq15scan10127jpg	2	2	2	2

ouq15scan10128jpg	3	3	3.5	3.5
ouq15scan10129jpg	2.5	2.5	2.5	2.5
ouq15scan10131jpg	5.5	5.5	4.5	5.5
ouq15scan10132jpg	4.5	4.5	4.5	4.5
ouq15scan10133jpg	1.5	1.5	1.5	1.5
ouq15scan10134jpg	0.5	0.5	0.5	0.5
ouq15scan10135jpg	0.5	1.5	1.5	1.5
ouq15scan10136jpg	3	3	4	4
ouq15scan10137jpg	1.5	1.5	1.5	1.5
ouq15scan10138jpg	4.5	4.5	5	5
ouq15scan10139jpg	3	3	3	3
ouq15scan10140jpg	3.5	3.5	4	4
ouq15scan10141jpg	2.5	2.5	2.5	2.5
ouq15scan10142jpg	2	1	1.5	1.5
ouq15scan10143jpg	6	6	5	6
ouq15scan10144jpg	3.5	3	3	3
ouq15scan10145jpg	5.5	5.5	4.5	5.5
ouq15scan10146jpg	4.5	4.5	4.5	4.5
ouq15scan10147jpg	5	4.5	3	4.5
ouq15scan10148jpg	1.5	1.5	2	2
ouq15scan10149jpg	2	2	2	2
ouq15scan10150jpg	2	2	2	2
ouq15scan10151jpg	1	1	1	1
ouq15scan10152jpg	2	3	2.5	3
ouq15scan10153jpg	3	2	2.5	2.5
ouq15scan10154jpg	3	3	2	3

ouq15scan10155jpg	0.5	0.5	0.5	0.5
ouq15scan10156jpg	3	3	3.5	3.5
ouq15scan10157jpg	1	1.5	1.5	1.5
ouq15scan10158jpg	2	2	2	2
ouq15scan10159jpg	3.5	3	3	3
ouq15scan10160jpg	2	2	2	2
ouq15scan10161jpg	3	3	3	3
ouq15scan10162jpg	5	4.5	4	4.5
ouq15scan10163jpg	1	1	1	1
ouq15scan10164jpg	3.5	3.5	3.5	3.5
ouq15scan10165jpg	5.5	5.5	5	5.5
ouq15scan10166jpg	4	4	3	4
ouq15scan10167jpg	3	2.5	3	3
ouq15scan10168jpg	5	5	6	6
ouq15scan10169jpg	3.5	3	2.5	3
ouq15scan10170jpg	2	2	2.5	2.5
ouq15scan10171jpg	1.5	1	1	1
ouq15scan10172jpg	4	4	4	4
ouq15scan10173jpg	4.5	4.5	4.5	4.5
ouq15scan10174jpg	5	5	5.5	5.5
ouq15scan10175jpg	7	7	5	7
ouq15scan10176jpg	4.5	4	4	4
ouq15scan10177jpg	1	1	0.5	1
ouq15scan10178jpg	5	4.5	4.5	4.5
ouq15scan10179jpg	4.5	4.5	4.5	4.5
ouq15scan10180jpg	3	3	3	3

ouq15scan10181jpg	3	3	2.5	3
ouq15scan10182jpg	1.5	1	1.5	1.5
ouq15scan10183jpg	4	4.5	3.5	4.5
ouq15scan10184jpg	5.5	5.5	3.5	5.5
ouq15scan10185jpg	0.5	1	1	1
ouq15scan10186jpg	5.5	5.5	4.5	5.5
ouq15scan10187jpg	3.5	3	3	3
ouq15scan10188jpg	2.5	2.5	3	3
ouq15scan10189jpg	0	0.5	0.5	0.5
ouq15scan10190jpg	2.5	2.5	2	2.5
ouq15scan10191	1.5	1.5	1.5	1.5
ouq15scan10192	1	1	1	1
ouq15scan10193	3.5	3.5	3.5	3.5
ouq15scan10194	4.5	4.5	4.5	4.5
ouq15scan10196	3.5	3.5	4	4
ouq15scan10197	4.5	4.5	4.5	4.5
ouq15scan10198	5	4	3.5	4
ouq15scan10199	6	6	5	6

Appendix C: Results for Corpus B Prior to Tool Modification

Student Diagram	Human Mark	Score(Best)
M358_Exam_06_pgt_01	1	1
M358_Exam_06_pgt_02	2.5	2
M358_Exam_06_pgt_03	0.5	1
M358_Exam_06_pgt_04	2	1
M358_Exam_06_pgt_05	1.5	1.5
M358_Exam_06_pgt_06	3	2.5
M358_Exam_06_pgt_07	2	2.5
M358_Exam_06_pgt_08	2.5	2.5
M358_Exam_06_pgt_09	2.5	2
M358_Exam_06_pgt_10	2.5	2
M358_Exam_06_pgt_11	3.5	3
M358_Exam_06_pgt_12	3.5	2.5
M358_Exam_06_pgt_13	2	1.5
M358_Exam_06_pgt_14	0.5	0
M358_Exam_06_pgt_15	1	0
M358_Exam_06_pgt_16	4	4.5
M358_Exam_06_pgt_18	4.5	4.5
M358_Exam_06_pgt_19	3	2
M358_Exam_06_pgt_20	2	2
M358_Exam_06_pgt_21	2.5	2
M358_Exam_06_pgt_22	3	2.5
M358_Exam_06_pgt_23	2.5	2.5
M358_Exam_06_pgt_24	3.5	2.5

M358_Exam_06_pgt_25	2.5	1.5
M358_Exam_06_pgt_26	4.5	3.5
M358_Exam_06_pgt_27	0	0
M358_Exam_06_pgt_29	3.5	3.5
M358_Exam_06_pgt_30	3.5	2.5
M358_Exam_06_pgt_31	2.5	2
M358_Exam_06_pgt_32	2	2

Appendix D: Specimen Solution Diagrams

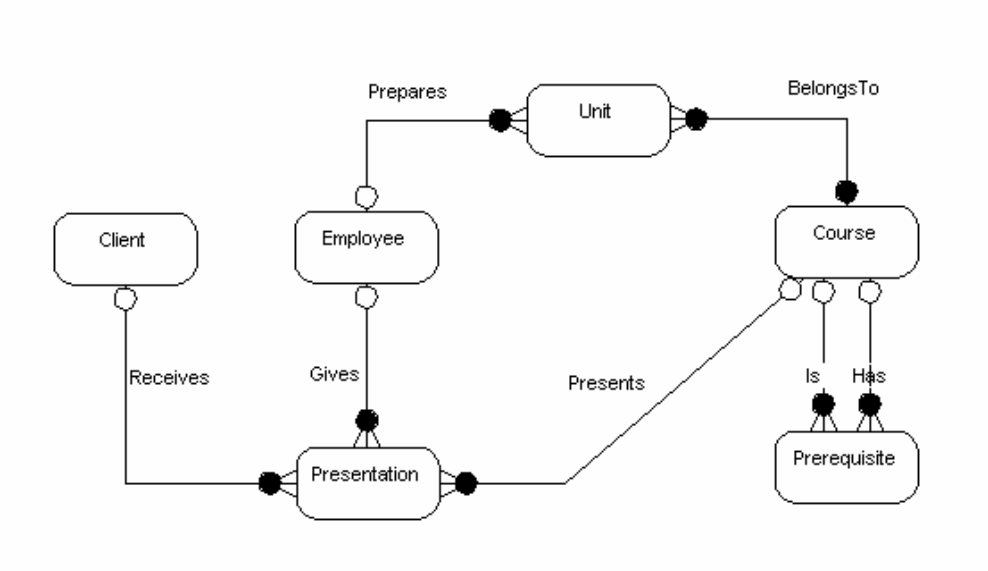

Figure: Specimen solution 0 for Corpus A

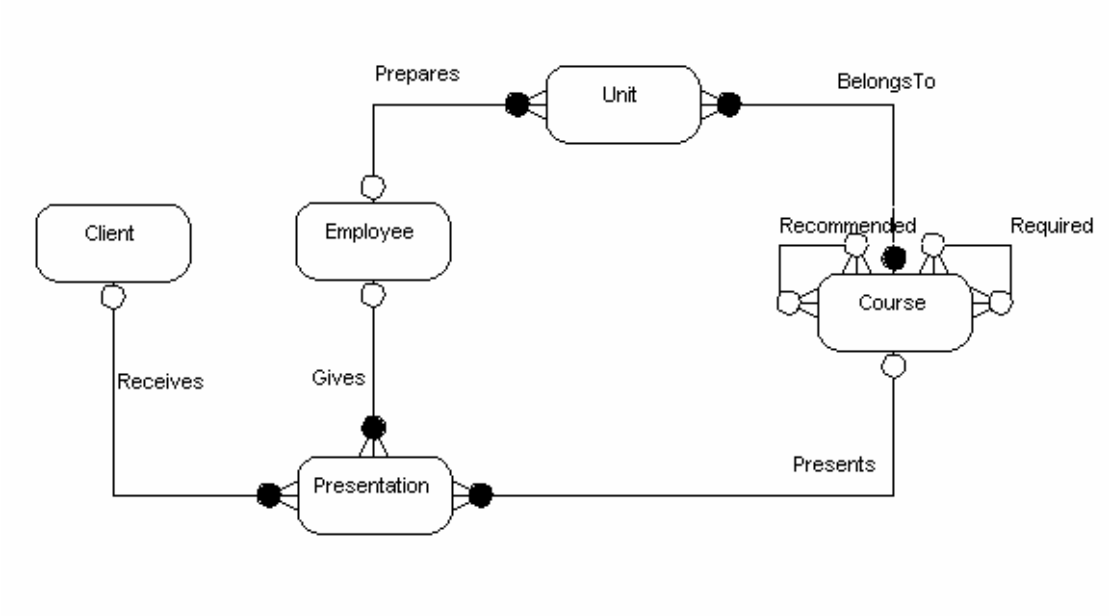

Figure: Specimen solution 1 for Corpus A

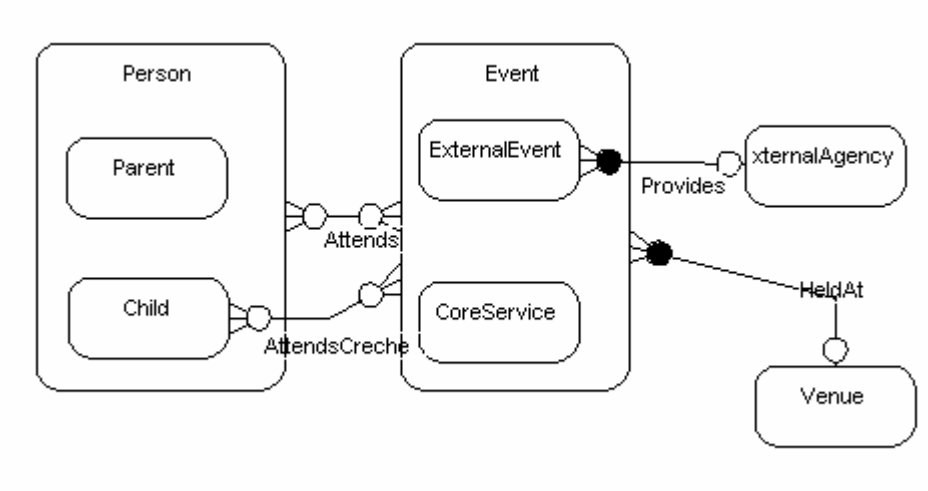

Figure: Specimen solution for Corpus B

Appendix E: Code to Correct for Incorrect Relationship Matching for Corpus B Diagrams

```
// © COPYRIGHT IBM CORP. 2007
// New instance variables
  ERDiagram myClone;
  boolean isClone = false;


// © COPYRIGHT IBM CORP. 2007
// New method
  private double performSubstituteMarking(ERDiagram other, int[][]
bestSubTypeMatch){

    myClone = (ERDiagramImpl)this.clone();
    myClone.setIsClone();

    Vector subs = myClone.getSubTypes();

    int[] missing = new int[subs.size()];
    for(int i=0; i<missing.length; i++){
      missing[i] = 0;
    }

    for(int i=0; i<missing.length; i++){
      for(int j=0; j<bestSubTypeMatch.length; j++){
        if(bestSubTypeMatch[j][0]==i){
          missing[i] = 1;
        }
      }

      Entity parent;
      Entity child;
      Association assoc;
      SubType mySubType;
      Vector rels = (Vector)myClone.getRelationships().clone();
      myClone.setRelationships(rels);
      myClone.setMarkScheme(this.getMarkScheme());
      Relationship relates;
      Relationship replace;
      Association compareNew = null;

      for(int j=0; j<missing.length; j++){
        if(missing[j]==0){
          mySubType = (SubType)subs.elementAt(j);
          parent = mySubType.getAssoc().getEntity(1);
          child = mySubType.getAssoc().getEntity(0);
```

```
        for(int k=0; k<rels.size(); k++){
          relates = (Relationship)rels.elementAt(k);
          Entity compare0 = relates.getAssociation().getEntity(0);
          Entity compare1 = relates.getAssociation().getEntity(1);
          if(compare0 == child||compare1 == child){
            if(compare0 == child){
              compareNew = new AssociationImpl(parent, compare1, 0);
            }
            if(compare1 == child){
              compareNew = new AssociationImpl(compare0, parent, 0);
            }
            Name nameNew = relates.getName();
            Degree degreeNew = relates.getDegree();
            Participation part0 = relates.getParticipation(0);
            Participation part1 = relates.getParticipation(1);
            replace = new RelationshipImpl(nameNew, compareNew, degreeNew, part0,
part1, 0);
            rels.set(k, replace);
          }
        }
      }
    }
   }
   double score = myClone.performMarking(other);
   System.out.println("Alt Marking Score = " + score);
   return score;
  }

// © COPYRIGHT IBM CORP. 2007
  public void setIsClone(){
    isClone = true;
  }

// © COPYRIGHT IBM CORP. 2007
  public void setRelationships(Vector v){
    relationships = v;
  }




   // Changes to performMarking() method
   // Now compute score
   MarkingCalculator mc = new MarkingCalculatorImpl(markScheme, this);
   score = mc.score(bestEntityMatch, bestRelMatch, bestSubTypeMatch, other);

   double score1 = 0;
   if(this.getSubTypes().size()>0&&!this.isClone){
     System.out.println("Start Substitute Marking");
     score1 = this.performSubstituteMarking(other, bestSubTypeMatch);
```

```
    System.out.println("Finish Substitute Marking");

  }
  markingPerformed = true;

  if(score1 > score){
    score = score1;
  }
  return score;
}
```

Appendix F: Code implementing Entity Exact Name Matching Method for Corpus A Diagrams

```
// Update to performMarking(ERDiagram other) method of class ERDiagramImpl
// at row 645
   EntityNameCompare enc = new EntityNameCompare(this, other, bestRelMatch);
   bestRelMatch = enc.checkEntityNames();



package marker07update;
import java.util.*;

/**
 * Title: Class EntityNameCompare which implements the Entity Exact Name
 *          Matching process
 * Copyright:    © COPYRIGHT IBM CORP. 2008
 * Company:      IBM
 * @author       Martin Ball
 * @version 1.0
 */

public class EntityNameCompare {

  private int[][] bestRelMatch;
  private ERDiagram other;
  private ERDiagram solution;
  private Hashtable entityNameMatches;
  private HashSet hs;

  public EntityNameCompare(ERDiagram sol, ERDiagram ot, int[][] brm) {

    solution = sol;
    other = ot;
    bestRelMatch = brm;
    entityNameMatches = new Hashtable();
    hs = new HashSet();
    fillHashtable();
    Enumeration en = entityNameMatches.keys();
    while(en.hasMoreElements()){
      Entity et = (Entity)en.nextElement();
      Vector v = (Vector)entityNameMatches.get(et);
      for(int i=0; i<v.size(); i++){
        Entity ev = (Entity)v.get(i);
      }
    }
  }
```

```
private void fillHashtable(){

  Vector otherEntities = other.getEntities();
  Vector solutionEntities = solution.getEntities();

  for(int i=0; i<solutionEntities.size(); i++){

    Entity e = (Entity)solutionEntities.get(i);
    String s = e.getNameString();
    s = s.trim();
    s = s.toLowerCase();

    Enumeration en = otherEntities.elements();

    while(en.hasMoreElements()){

      Entity f = (Entity)en.nextElement();

      String t = f.getNameString();
      t = t.trim();
      t = t.toLowerCase();

      if(s.equalsIgnoreCase(t)){
        System.out.println("s = " + s + " t = " + t);
        addEntity(e, f);
      }
    }
  }
}



private void addEntity(Entity key, Entity value){
  hs.add(value);
  if(entityNameMatches.containsKey(key)){
    Vector v = (Vector)entityNameMatches.get(key);
    v.add(value);
  }
  else{
    Vector u = new Vector();
    u.add(value);
    entityNameMatches.put(key, u);
  }
}

public int[][] checkEntityNames(){

  Vector vsol = solution.getRelationships();
  Vector vother = other.getRelationships();
```

```
  for(int i=0; i<bestRelMatch.length; i++){
    if(bestRelMatch[i][0]>=0){
      System.out.println("Test 1 " + bestRelMatch[i][0]);
      Relationship rel = (Relationship)vsol.get(bestRelMatch[i][0]);
      Entity e0 = rel.getAssociation().getEntity(0);
      Entity e1 = rel.getAssociation().getEntity(1);
      Relationship oth = (Relationship)vother.get(i);
      Entity e2 = oth.getAssociation().getEntity(0);
      Entity e3 = oth.getAssociation().getEntity(1);
      if(entityNameMatches.containsKey(e0)||entityNameMatches.containsKey(e1)){
        System.out.println("Test 2 " + e0.toString() + "   " + e1.toString());
        if(!doRelationshipsMatch(e0, e1, oth)){
          bestRelMatch[i][0] = -1;
          bestRelMatch[i][1] = -1;
          bestRelMatch[i][2] = -1;
        }
      }
    }
  }
  return bestRelMatch;
}


private boolean doRelationshipsMatch(Entity e0, Entity e1, Relationship oth){
  boolean b = true;
  Entity e3 = oth.getAssociation().getEntity(0);
  Entity e4 = oth.getAssociation().getEntity(1);
  Vector v0 = new Vector();
  Vector v1 = new Vector();
  if(entityNameMatches.containsKey(e0)){
    v0 = (Vector)entityNameMatches.get(e0);
  }
  if(entityNameMatches.containsKey(e1)){
    v1 = (Vector)entityNameMatches.get(e1);
  }

  if(entityNameMatches.containsKey(e0)){
    System.out.println("Test 4 " + v0.contains(e3) + "   " + v0.contains(e4));
    if(!(v0.contains(e3)||v0.contains(e4))){
      b = false;
    }
  }

  if(entityNameMatches.containsKey(e1)){
    System.out.println("Test 5 " + v0.contains(e4) + "   " + v1.contains(e4));
    if(!(v1.contains(e3)||v1.contains(e4))){
      b = false;
    }
  }
  if(hs.contains(e3)){
```

```
      System.out.println("Test 7 " + v0.contains(e3) + "    " + v1.contains(e3));
      if(!(v0.contains(e3)||v1.contains(e3))){
        b = false;
      }
    }
    if(hs.contains(e4)){
      System.out.println("Test 8 " + v0.contains(e4) + "    " + v1.contains(e4));
      if(!(v0.contains(e4)||v1.contains(e4))){
        b = false;
      }
    }
    System.out.println("Test 6 " + b);
    return b;
  }
}
```

Appendix G: Results for Extended Corpus A Prior to Tool Modification

Note that these results were taken with parameter Synonym Check set to False.

Student Diagram	Human Mark	Score(0)	Score(1)	Score(Best)
ouq15scan10200	5	5	4.5	5
ouq15scan10201	4	5	4	5
ouq15scan10202	4	4	4	4
ouq15scan10203	2.5	2.5	3	3
ouq15scan10204	4	4	4	4
ouq15scan10205	2.5	3.5	3.5	3.5
ouq15scan10206	6	6	5	6
ouq15scan10207	6.5	6.5	4.5	6.5
ouq15scan10208	3.5	3.5	2.5	3.5
ouq15scan10209	1.5	1.5	1.5	1.5
ouq15scan10210	2.5	2.5	2	2.5
ouq15scan10211	5	5	5.5	5.5
ouq15scan10212	1.5	1.5	1.5	1.5
ouq15scan10213	5.5	5.5	4.5	5.5
ouq15scan10214	2.5	2.5	2.5	2.5
ouq15scan10215	2	1	1.5	1.5
ouq15scan10216	4	4	4	4
ouq15scan10217	3	3	3	3
ouq15scan10218	1.5	1.5	2	2
ouq15scan10219	0.5	0.5	0.5	0.5
ouq15scan10220	3	3	3	3
ouq15scan10221	2.5	2.5	2	2.5
ouq15scan10222	3.5	3.5	3.5	3.5
ouq15scan10223	4	3	3.5	3.5

ouq15scan10224	2.5	2.5	2.5	2.5
ouq15scan10225	4.5	3.5	3.5	3.5
ouq15scan10226	3.5	3.5	2.5	3.5
ouq15scan10227	2.5	2.5	1.5	2.5
ouq15scan10228	4.5	4.5	4.5	4.5
ouq15scan10229	4.5	4.5	4.5	4.5
ouq15scan10230	4	4	4	4
ouq15scan10231	3	3	3	3
ouq15scan10232	1	1	1.5	1.5
ouq15scan10233	1	3	3	3
ouq15scan10234	4.5	4.5	4	4.5
ouq15scan10235	7	6	5	6
ouq15scan10236	3	3	2.5	3
ouq15scan10237	5	5	5.5	5.5
ouq15scan10238	2.5	2.5	3	3
ouq15scan10239	7	7	5	7
ouq15scan10240	0.5	1	1	1
ouq15scan10241	5	5	5.5	5.5
ouq15scan10242	1	1	1	1
ouq15scan10243	4.5	4.5	3.5	4.5
ouq15scan10244	4	4	4	4
ouq15scan10245	4	4.5	3.5	4.5
ouq15scan10246	3	3	3.5	3.5
ouq15scan10247	1.5	1.5	1.5	1.5
ouq15scan10248	0.5	1	1	1
ouq15scan10249	3	3	4	4
ouq15scan10250	2	2	2.5	2.5
ouq15scan10251	3	3	3	3

ouq15scan10252	2	4	4	4
ouq15scan10253	4.5	4.5	4.5	4.5
ouq15scan10254	2	1.5	1.5	1.5
ouq15scan10255	3	3	3	3
ouq15scan10256	2	1.5	2	2
ouq15scan10257	5.5	4.5	4.5	4.5
ouq15scan10258	5	5	4.5	5
ouq15scan10259	3.5	3.5	2.5	3.5
ouq15scan10261	4	4	4	4
ouq15scan10263	2	2	1.5	2
ouq15scan10264	3.5	3.5	4	4
ouq15scan10265	3.5	3	3	3
ouq15scan10266	4.5	4.5	3.5	4.5
ouq15scan10267	2.5	3	2.5	3
ouq15scan10268	4.5	4.5	4.5	4.5
ouq15scan10269	5	5	5	5
ouq15scan10270	4	4	5	5
ouq15scan10271	2	2	2	2
ouq15scan10272	4	4	4	4
ouq15scan10273	6	4	5	5
ouq15scan10274	6.5	6.5	4.5	6.5
ouq15scan10275	1.5	1.5	1.5	1.5
ouq15scan10276	2	1.5	1.5	1.5
ouq15scan10277	3	4	4.5	4.5
ouq15scan10278	2	2	2	2
ouq15scan10279	4.5	4.5	4.5	4.5
ouq15scan10280	3	2.5	2	2.5
ouq15scan10281	4.5	4.5	5	5

ouq15scan10282	2	2	2	2
ouq15scan10283	4.5	4.5	5	5
ouq15scan10284	1	2	1.5	2
ouq15scan10285	2.5	3	3	3
ouq15scan10286	0	0	0	0
ouq15scan10287	4.5	4.5	2.5	4.5
ouq15scan10288	1	0	0	0
ouq15scan10289	1	1.5	1.5	1.5
ouq15scan10290	5	5	5	5
ouq15scan10291	3.5	3	3	3
ouq15scan10292	4.5	4.5	4.5	4.5
ouq15scan10293	2.5	2.5	2	2.5
ouq15scan10294	5.5	5.5	4.5	5.5
ouq15scan10295	5	5	4	5
ouq15scan10296	4.5	4.5	5	5
ouq15scan10297	3	3	4	4
ouq15scan10298	2	2.5	2.5	2.5
ouq15scan10299	2	2	2	2
ouq15scan10300	3.5	3.5	3.5	3.5
ouq15scan10301	7	7	5	7
ouq15scan10302	3	3	2.5	3
ouq15scan10303	2.5	2.5	2	2.5
ouq15scan10305	7	7	5	7
ouq15scan10306	5	5	5	5
ouq15scan10307	2	2	1.5	2
ouq15scan10308	1	2	2	2
ouq15scan10309	3.5	3.5	2.5	3.5
ouq15scan10310	2	2	2	2

ouq15scan10311	4.5	4.5	3.5	4.5
ouq15scan10312	3	3	3	3
ouq15scan10313	0.5	0.5	0.5	0.5
ouq15scan10314	4.5	4.5	4.5	4.5
ouq15scan10315	3.5	4.5	5	5
ouq15scan10316	5.5	5.5	4.5	5.5
ouq15scan10317	2	2	1.5	2
ouq15scan10318	1.5	2	2	2
ouq15scan10319	1	1	1	1
ouq15scan10320	3.5	4	3	4
ouq15scan10321	4.5	4.5	4	4.5
ouq15scan10322	2	1.5	1	1.5
ouq15scan10323	6.5	4.5	5.5	5.5
ouq15scan10324	3.5	3.5	3.5	3.5
ouq15scan10325	3.5	3.5	3.5	3.5
ouq15scan10326	3	3	3	3
ouq15scan10327	1.5	1.5	1.5	1.5
ouq15scan10328	3	3	3.5	3.5
ouq15scan10329	3	3	3	3
ouq15scan10330	3	3	3	3
ouq15scan10331	2	2	2	2
ouq15scan10332	2.5	2.5	2.5	2.5
ouq15scan10333	6	6	4	6
ouq15scan10334	6.5	4.5	5.5	5.5
ouq15scan10335	2	2	2	2
ouq15scan10336	6	4	5	5
ouq15scan10337	5.5	4.5	6.5	6.5
ouq15scan10339	3.5	3.5	4	4

ouq15scan10340	1.5	2	2	2
ouq15scan10341	4	3	3.5	3.5
ouq15scan10342	3.5	3.5	3.5	3.5
ouq15scan10343	1.5	1.5	1.5	1.5
ouq15scan10344	1.5	1.5	1.5	1.5
ouq15scan10345	2	3	2.5	3
ouq15scan10346	4.5	4	3	4
ouq15scan10347	2.5	2.5	3.5	3.5
ouq15scan10348	1.5	2	2	2
ouq15scan10349	4	4	4.5	4.5
ouq15scan10350	5	5	5.5	5.5
ouq15scan10351	4.5	4.5	5.5	5.5
ouq15scan10352	3.5	3.5	4	4
ouq15scan10354	3.5	3.5	4	4
ouq15scan10355	4	4	4.5	4.5
ouq15scan10356	2.5	2.5	2	2.5
ouq15scan10357	3	3	3	3
ouq15scan10358	3.5	3	3	3
ouq15scan10359	5	5	5	5
ouq15scan10360	6	6	5	6
ouq15scan10361	4.5	4.5	4.5	4.5
ouq15scan10362	2.5	2.5	2.5	2.5
ouq15scan10363	2	1.5	1.5	1.5
ouq15scan10364	5	5	4	5
ouq15scan10365	5	5	4	5
ouq15scan10366	2	2	2.5	2.5
ouq15scan10367	4.5	4.5	3	4.5
ouq15scan10368	5.5	5	5	5

ouq15scan10369	3.5	3	2.5	3
ouq15scan10370	6	4.5	4.5	4.5
ouq15scan10371	4.5	4.5	4.5	4.5
ouq15scan10372	2.5	1.5	1.5	1.5
ouq15scan10373	7	7	5	7
ouq15scan10374	3.5	3.5	2.5	3.5
ouq15scan10375	7	7	5	7
ouq15scan10376	2	2	2	2
ouq15scan10377	3.5	4	4	4
ouq15scan10378	4	4	4	4
ouq15scan10379	4.5	4.5	4.5	4.5
ouq15scan10380	1.5	1.5	1.5	1.5
ouq15scan10381	5	4	4	4
ouq15scan10382	4	4	4	4
ouq15scan10383	2	2	2	2
ouq15scan10384	2	2	2.5	2.5
ouq15scan10385	2.5	2.5	2.5	2.5
ouq15scan10386	3	3	2.5	3
ouq15scan10387	5.5	5	5	5
ouq15scan10388	0.5	0.5	0.5	0.5
ouq15scan10389	3.5	3.5	4	4
ouq15scan10390	2.5	2.5	2.5	2.5
ouq15scan10391	3	3	2.5	3
ouq15scan10392	2	2.5	3	3
ouq15scan10393	0.5	0.5	0.5	0.5
ouq15scan10394	4.5	4.5	3.5	4.5
ouq15scan10395	2	2	2	2
ouq15scan10396	4	5	4.5	5

ouq15scan10397	1	1	1	1
ouq15scan10398	3	3	2.5	3
ouq15scan10399	2.5	2.5	2.5	2.5