Learning Analytics for Academic Writing through Automatic Identification of Meta-discourse

Thesis

LEARNING ANALYTICS FOR ACADEMIC WRITING THROUGH AUTOMATIC IDENTIFICATION OF META-DISCOURSE

Duygu Bektik (née Simsek)

BA & MA (with 1st class Honours Degree)
in Computer and Instructional Technology Teacher Education
Bilkent University, 2011

MSc. (with Distinction)
in Software Engineering
University of Southampton, 2012

THESIS SUBMITTED IN PARTIAL FULFILMENT OF THE
REQUIREMENTS FOR THE DEGREE OF
DOCTOR OF PHILOSOPHY

KNOWLEDGE MEDIA INSTITUTE
THE OPEN UNIVERSITY

SEPTEMBER 2016

Duygu Bektik: Learning Analytics for Academic Writing Through Automatic Identification of Meta-Discourse.

Submitted in September 2016. Examined in December 2016.

Supervisors:

Prof. Denise Whitelock

Dr. Rebecca Ferguson

Dr. Anna De Liddo

External Supervisor:

Prof. Simon Buckingham Shum, University of Technology Sydney

Examiners:

Dr. Paul Mulholland, The Open University UK

Dr. Gary Wills, University of Southampton UK

ABSTRACT

Effective written communication is an essential skill which promotes educational success for undergraduates. Argumentation is a key requirement of successful writing, which is the most common genre that undergraduates have to write particularly in the social sciences. Therefore, when assessing student writing academic tutors look for students' ability to present and pursue well-reasoned and strong arguments through scholarly argumentation, which is articulated by meta-discourse.

Today, there are some natural language processing systems which automatically detect authors' rhetorical moves in scholarly texts. Hence, when assessing their students' essays, educators could benefit from the available automated textual analysis which can detect meta-discourse. However, previous work has not shown whether these technologies can be used to analyse student writing reliably. The aim of this thesis therefore has been to understand how automated analysis of meta-discourse in student writing can be used to support tutors' essay assessment practices. This thesis evaluates a particular language analysis tool, the Xerox Incremental Parser (XIP) as an exemplar of this type of automated technology.

The studies presented in this thesis investigates how tutors define the quality of undergraduate writing and suggests key elements that make for good quality student writing in the social sciences, where XIP seems to work best. This thesis also sets out the changes that needs to be made to the XIP and proposes in what ways its output can be delivered to tutors so that they make use of this output to give feedback on student essays. The findings reported also show problems that academic tutors experience in essay assessment, which potentially could be solved by automated support. However, tutors have preconceptions about the use of automated support. The study revealed that tutors want to

be assured that they retain the ‘power’ themselves in any decision of using automated support to overcome these preconceptions.

This thesis is dedicated to my beloved husband…

ACKNOWLEDGEMENTS

Writing this thesis has been a long learning journey for which I owe several people a debt of gratitude.

First and foremost, I would like to express my sincerest gratitude to my supervisors, Prof. Denise Whitelock, Dr. Rebecca Ferguson, Dr. Anna De Liddo and my external supervisor Prof. Simon Buckingham Shum, for their invaluable guidance, and constant prompt feedback. I would also like to acknowledge Dr. Ágnes Sándor, from Xerox Research Centre Europe, for her collaboration, enthusiasm and willingness to brainstorm and discuss my research ideas and findings.

My deep appreciation and very special thanks go out to Prof. Kevin McConway, for his invaluable statistics advice, mentoring and feedback on all the quantitative aspects of my research; and to Dr. Jackie Tuck, whose excellent help during data collection has made an invaluable contribution towards my PhD. I would also like to thank Dr. Clara Mancini, my third party monitor, for her willingness to help me with anything, any time I needed it.

I gratefully acknowledge the financial support and funding received towards my PhD from the Wolfson Open Science Laboratory, and the Knowledge Media Institute. I would also like to thank and appreciate the Institute of Educational Technology for providing an additional work environment for my studies.

I would like to thank my friends who have provided a continual reminder that there are other joys in life besides writing up a thesis. And of course a big thank you to Gillian Çağlayan, who generously gave up her time for helping me enormously with the mammoth task of proofreading of this thesis, thus improving it.

I believe that I am tremendously fortunate to have the outstanding support of my family, and am deeply appreciative to all of them, for their encouragement, and prayers. Special thanks to my mother, Belgin Özkul, who has supported me throughout this long journey,

and for all of the sacrifices she has made on my behalf. In particular, I have been extremely fortunate in my life to have my grandmother, Nazife Özkul, who has shown me unconditional love and support in every step of my life, which holds an enormous amount of meaning to me. My dear granny, from now on you no longer need to worry that I will go crazy one day from working too hard; I am done with all my studies.

And finally, but by no means least, I am grateful for the love and tolerance of Mustafa Bektik, the man who has made all the difference in my life. Without his patience and sacrifice, I could not have completed this thesis. To Mustafa, my beloved husband, the cornerstone of my life and the making of me. Thank you for your encouragement, patience, understanding, and support as I snatched moments around family life to write this thesis. I am ready now to live every moment to the fullest!

SHORT CONTENTS

CONTENTS

GLOSSARY

The Open University (OU) is a distance learning and research university founded by Royal Charter in the United Kingdom. It has an open entry policy, and nearly all of its courses have no entry requirements, i.e. students' previous academic achievements are not taken into account for entry to most undergraduate courses. The OU's undergraduate students study off-campus, undertaking e-learning activities, with the flexibility of deciding when and where to study day-to-day. This glossary is created in order to explain specific terms and definitions used by the OU when referring to courses and job roles.

Course: A programme of study leading to a degree or other qualification.

Module: For each stage of a course, students take a range of modules, which normally involve a specific set of classes on a particular topic, and a specific set of accompanying assignments to complete. Each module has its own code and title, e.g. S288 Practical Science ('S' stands for Science and '2' represents a level 2 module).

1st/2nd/3rd levels: At the OU, a level is equivalent to a 1st/2nd/3rd-year course at a bricks and mortar university.

Levels: The modules students take each have a 'level', which tells the relative demand, complexity and depth of the work required. When students complete a Level 1 module, the credit they gain for it is Level 1 credit; a Level 2 module gives Level 2 credit, and so on.

Credit system: The credit system used by the OU is aligned to the national Credit Accumulation and Transfer Scheme (CATS). The CATS system helps students to move the credits they accumulate from one institution to another. If they have completed some previous study at another institution, they may be able to count it towards their OU qualification. The scheme equates one credit (or credit point) with 10 hours of notional learning time (the time, on average, a learner takes to achieve the specified learning outcomes). A module allocated 30 credits or 'CATS points' should require students to

commit approximately 300 hours of work to achieving the learning outcomes for the module at a particular level. OU modules carry credits normally ranging from 30 to 60 credits, and will be awarded if students satisfactorily complete a module.

Module team: Module teams oversee the creation and delivery of OU modules. The OU has two types of module team: production and presentation. The production team designs module pedagogy and content, and produces learning materials prior to presentation. The presentation module team maintains a module during its presentation to students. The curriculum manager works collaboratively with the module team chair to ensure that students have a consistent and high quality experience of the module in each presentation. This includes ensuring the learning materials remain up-to-date and accurate, and maintaining an engaging website for students, managing budgets and contributing to the management of end-of-module examination process.

Module team chairs: The OU module team chairs have the responsibility of providing academic leadership to ensure the coherence, cohesion and quality of the learning experience offered by the module. They have overall responsibility for the work of the module team, and the setting and maintenance of academic standards. Their specific responsibilities are academic leadership, quality assurance and enhancement, research and scholarship, curriculum development, management and administration, and module development and delivery.

Associate lecturers (ALs): Associate lecturers are teaching professionals employed to teach OU students. They are responsible for ensuring the provision of high quality teaching and learning, and support students by making contact at key points in their study, marking and providing feedback on assignments, helping students understand the module material, and helping them prepare for an examination or end of module assessment.

End-of-module assessment (EMA): OU modules usually have two assessment components: continuous assessment (assignments completed during the module) and an examination or other examinable piece of work such as a project, portfolio or dissertation.

These pieces of examinable work are usually referred to as end-of-module assessments or EMAs.

Tutor-marked assignment (TMA): Continuous assessment can be made up of TMAs or CMAs (computer marked assignments). TMAs are usually essays or short-answer questions.

Assessment scale: Although some modules use other scales for marking individual assignments, and those will be provided to the students in their assignment material, the overall score for an assignment is recorded and reported using the University Scale, as explained below.

Band	University scale score	Performance standard
A	85-100	Pass 1
B	70-84	Pass 2
C	55-69	Pass 3
D	40-54	Pass 4
E	30-39	Bare fail
F	15-29	Fail
G	0-14	Bad fail

Coordination meetings: On most modules with an EMA, a coordination meeting is held where markers meet with the module team to agree a common interpretation of the marking guide prior to the marking of examinable work. The coordination may be a face-to-face meeting held on the university's main campus, or may take place online/via telephone. Markers are provided with the marking guide and asked to mark a selection of sample EMAs ahead of the exercise.

[This page intentionally left blank.]

PREFACE

The work presented in this thesis has led to the following publications, listed in chronological order.

- Simsek, Duygu; Buckingham Shum, Simon; Sandor, Agnes; De Liddo, Anna and Ferguson, Rebecca (2013). XIP Dashboard: visual analytics from automated rhetorical parsing of scientific *metadiscourse*. In: 1st International Workshop on Discourse-Centric Learning Analytics, 8 Apr 2013, Leuven, Belgium.
- Taibi, Davide; Sandor, Agnes; Simsek, Duygu; Buckingham Shum, Simon; De Liddo, Anna and Ferguson, Rebecca (2013). Visualizing the LAK/EDM literature using combined concept and rhetorical sentence extraction. In: Proceedings of the LAK Data Challenge, 3rd Int. Conf. on Learning Analytics and Knowledge (LAK '13), 8-12 April 2013, Leuven, Belgium.
- Simsek, Duygu; Buckingham Shum, Simon; De Liddo, Anna; Ferguson, Rebecca and Sándor, Ágnes (2014). Visual analytics of academic writing. In: Proceedings of the Fourth International Conference on Learning Analytics and Knowledge - LAK '14, pp. 265–266.
- Simsek, Duygu; Sandor, Agnes; Buckingham Shum, Simon; Ferguson, Rebecca; De Liddo, Anna and Whitelock, Denise (2015). Correlations between automated rhetorical analysis and tutors' grades on student essays. In: Proceedings of the Fifth International Conference on Learning Analytics and Knowledge, ACM, pp. 355–359.
- Buckingham Shum, Simon; Knight, Simon; McNamara, Danielle; Allen, Laura; Bektik, Duygu; & Crossley, Scott (2016). Critical perspectives on writing

analytics. In: Proceedings of the Sixth International Conference on Learning Analytics & Knowledge (pp. 481-483). ACM.

The data in the study, as given in Chapter 5.3, come from the British Academic Written English (BAWE) corpus, which was developed at the Universities of Warwick, Reading and Oxford Brookes under the directorship of Hilary Nesi and Sheena Gardner (formerly of the Centre for Applied Linguistics [previously called CELTE], Warwick), Paul Thompson (formerly of the Department of Applied Linguistics, Reading) and Paul Wickens (Westminster Institute of Education, Oxford Brookes), with funding from the ESRC (RES-000-23-0800).

The design of the rest of the studies presented in this thesis were examined by The Open University's Human Research Ethics Committee (HREC), and approved as follows:

- HREC/2013/1510/Simsek
- HREC/2014/1510/Simsek-2
- HREC/2014/1510/Simsek-3

Pseudonyms are used for all the Open University module codes and names to ensure the anonymity of participants. The descriptions of the interviewees provided in chapter 6.2 were written in a way that ensured none of the participants were easily identifiable. These descriptions were all checked with the participants themselves.

1

INTRODUCTION

1.1 Writing

The complexity of writing tasks is varied. Depending on the type of writing activity, it can be extremely difficult or quite straightforward. It is fair to say that on a typical day most students and academics perform some sort of writing, some of which is trivial. Although everyday writing tasks often seem to be effortless, such as jotting down a to-do list for the day, or leaving a note to a milkman, when it comes to writing an essay for a university assignment it takes much longer to put ideas down on paper, even though using information technology allows us to write faster than writing by hand.

Typing and writing are not necessarily the same activities, however. They are carried out in different places and with different equipment. Computers produce text in response to dictation, and correct spelling automatically most of the time. Text written by hand using pen and paper can be re-arranged on the page easily and can incorporate little doodles and diagrams. Without getting into the discussion of whether 'the pen is still mightier than the keyboard' (Chemin, 2014), when this thesis talks about writing, it actually means talking about the act of writing using a computer.

Writing for an assignment requires more effort than jotting down a to-do list. An obvious reason is that every day scribbling does not require structured writing, sentences or paragraphs, unlike a university assignment (Sharples, 1999). Most everyday handwriting is unstructured, short, and may be listed in bullets for non-critical readers (such as writing

reminders on a post-it note: 1) money to milkman, 2) grocery shopping, 3) dentist appointment at 3pm, etc.). Writing such notes requires authors to structure their thoughts, rather than their writing, to make sure everything they want to say or remember is recorded on the paper.

University assignments require more complex structured writing prepared for a critical reader. Such writing is composed for 'knowledge telling', restating back what has been already known such as definitions, facts, and theories; and/or for 'knowledge transforming', meaning retelling the knowledge with a new perspective and with a critical eye (Grabe & Kaplan, 1996).

When the knowledge is created and transformed, it requires not only structuring one's own thoughts but also configuring for the critical reader so that s/he can grasp the meaning. As Vygotsky (1988) observed, writing is not simply writing what has been already known (knowledge telling), but it requires pre-processing what we intend to write through inner speech:

> It is no surprise that written speech is the polar opposite of oral speech. The situation that is clear to the interlocutors in oral speech, and the potential for expressive intonation, mimic and gesture, is absent in written speech. The potential for abbreviation is excluded from the outset. Understanding must be produced through words and their proper combination. Written speech facilitates speech as a complex activity. This underlies the use of the rough draft. The path from the rough to the final draft is a complex activity. However, even without the rough draft, the process of reflecting on one's work in written speech is extremely powerful. Frequently, we say what we will write to ourselves before we write. What we have here is a rough draft in thought...this rough draft that is constructed in thought as part of written speech is inner speech. Inner speech acts as an internal rough draft in oral as well as in written speech.

"Writing is analytic, requiring evaluation and problem solving, yet it is also a synthetic, productive process" (Sharples, 1999, p. 10). Although most people know the structure of a thank you letter, few people can turn out a university assignment, short story, a technical report without preparing the structure in advance (Sharples, 1999). Texts longer than a

couple of paragraphs generally conform to an overall structure that frames the style and content of the text and organises the expectations of the reader.

The difficulty of writing a university assignment could be explained as more than the need for a macro-structure; the writing of such essays is usually a thought-provoking activity that requires particular skills of critical thinking and argumentation which are not 'a student's mother tongue' (Sommers & Saltz, 2004). Accepting Toulmin's (2003) definition, within the context of this thesis, argument can be defined as an assertion or a claim and its justification (Toulmin, 2003). The term argumentation can be defined as the process of arguing, giving reasons where claims are justified and evidenced through written text. Students, especially those in their first year at university, are unused to this form of writing, and most of them see themselves as novices (Sommers & Saltz, 2004). In order to understand how computers can help with academic writing, it is essential to look at both the literature on academic writing in general as well as what academic writing actually means, particularly in higher education, and what students are required to do to deliver a good essay. This is discussed in the first part of the literature review, in next chapter.

1.2 Undergraduate writing

One of the key requirements of good academic writing in undergraduate higher education courses is that students must develop a critical mind, and learn how to construct sound arguments in their discipline. When assessing student essays, educators look for students' ability to present and pursue well-reasoned and strong arguments, and for their ability to use examples, and evidence for and against a particular position.

Research investigating the fundamental features of good student writing has found that argumentation is a key requirement. "Argumentation is a key requirement of the essay, which is the most common genre that students have to write" (Wingate, 2012, p. 145).

However, since novice student writers join higher education with partial or incorrect conceptions about argumentation (Wingate, 2012), they are not familiar with what they are expected to produce and they have difficulty in constructing argumentative pieces. Undergraduate students struggle with argumentation: they are either unaware that they are expected to develop an argument in their essays, or have difficulty in arguing (Bacha, 2010; Davies, 2008), often because they have learned different concepts of argument at secondary school (Andrews, 1995).

Additionally, between students and their academic tutors, there are mismatched views regarding how student academic writing should be carried out and what good student academic writing looks like. The problem is not only that novice student writers come from relatively different backgrounds of school culture, but also because argumentation is often not adequately explained by their academic tutors, who often only vaguely understand how argumentation is realised in writing in different disciplines and therefore struggle to provide effective feedback which would prompt good examples of argumentation (Lea & Street, 1998; Mutch, 2003).

1.3 Role of meta-discourse in argumentation

Argumentation is articulated by *meta-discourse*. Meta-discourse refers to the features of text that provide linguistic cues which engage the readers, and explicitly convey the authors' intended meaning, expressing their viewpoint, argument and claim, and signaling their stance (Hyland, 2005). Rather than simply defining meta-discourse as 'discourse about discourse', Hyland (2005) defined the concept of 'meta-discourse' as an important element of the document, that is not only used to organise ideas but also to relate to readers. It is an umbrella term that helps to relate the text to its context, which glues the important parts of a text together but, more significantly, it helps readers to understand existing knowledge and strategies used by other members (authors/researchers) of the subject area, as well as the writer's stance towards these.

One of the aims of writing is to convince readers of the validity of claims and arguments (de Waard *et al.*, 2009). Therefore, authors put forward their claims to persuade and engage readers that their argument is valid. Since meta-discourse signals the problem, claim, argument, and evidence within writing, it allows readers to clarify the author's intention, to make meaning of a text (Williams, 2007).

When assessing student writing, academic tutors, as critical readers, look for students' ability to present and pursue well-reasoned and strong arguments through scholarly argumentation. Academic tutors will therefore necessarily be examining meta-discourse in students' writing as signals of the intellectual moves that make their reasoning visible.

1.4 Automatic identification of meta-discourse

Natural language processing (NLP) is the automatic processing of human language, natural language, such as English, rather than a specialised artificial computer language. "NLP is the application of computational methods for the purpose of analysing language-related characteristics of electronic files of text or speech" (Shermis & Burstein, 2013, p. 56). Today, some natural language processing systems exist which automatically detect authors' rhetorical moves in scholarly/academic texts. One approach to automatic identification of rhetorical moves, meta-discourse, is 'Argumentative Zoning' (Teufel, 1999), and another approach is 'Xerox Incremental Parser' (Aït-Mokhtar, Chanod, & Roux, 2002), both explained in detail in the next chapter.

The archaic definition of rhetoric is the art and study of the use of language with persuasive effect in any given field (Dawson, 1998); the art of trickery, a way of masquerading and obscuring information (Maynard, 1998). A more contemporary definition of rhetoric refers to the skill to analyse, evaluate and employ writing strategies in order to respond to the audience and being aware of one's own ideological stance and the audience's stance (Cook, 2002). Rhetorical 'move' refers to "a discoursal or rhetorical unit

that performs a coherent communicative function in a written or spoken discourse" (Swales, 2004, p. 228).

When assessing their students' essays, educators could benefit from the available automated textual analysis which can detect meta-discourse. In this way, academic tutors could also overcome the issue of providing effective practical examples of what argumentation should look like with readily available automatic machine output. However, the issue of whether these technologies can be used to analyse student writing reliably has not yet been adequately investigated.

This thesis investigates whether computational language technologies can automatically identify the attributes of good academic writing in undergraduate student writing and, if this proves possible, how best to feed back an actionable analysis in order to support educators in their essay assessment processes. In particular, it examines what makes good quality student writing, whether student writing across disciplines can be analysed automatically in order to detect the presence of good quality writing, and how automated support should be integrated into tutors' essay assessment practices.

1.4.1 Automated Support

Assessing written texts is a labour-intensive process for academic tutors. Marking and giving detailed feedback, and commenting on essays can be time-consuming, which was the reason for Page (1966) to come up with an idea of using computers to analyse essays automatically back in late 1960s, and to support tutors who spend so much time marking papers and are then burdened with hours of grading them.

Although automated support seems to be useful to overcome the problems stated above, there has been an ongoing tension between the writing teachers, researchers, academic tutors and markers on the one side and the developers of such automated technologies on the other regarding the use of automated essay evaluation. "There is an inherent suspicion that technology can corrupt the essence of a fundamentally human activity" (Elliot &

Williamson, 2013). Since many tutors see automated technology as a threat instead of support, this tension has often appeared in academic literature. For instance, Carl Whithaus, a writing teacher and researcher at the University of California, Davis, opened up his foreword for the *Handbook of Automated Essay Evaluation* with the following:

> When I was invited to write this foreword, I was both delighted and terrified. I was delighted because two of the leading researchers of Automated Essay Evaluation (AEE) software were asking a writing teacher and researcher to weigh in on a collection of essays about "the state of the art" in their field. I was terrified because the drumbeat of criticism about AEE, or "robo-grading" as its critics like to call it, had jumped from relatively obscure academic forums… I was worried because writing this foreword would very publicly place me in the camp of being a proponent of AEE. No matter how guarded and nuanced my position in favour of AEE was, it was certain to be criticised by a wide range of my colleagues and friends in writing studies. I did wonder whether I should write this foreword or not (Shermis & Burstein, 2013, p. vii).

The tension is originated by the awareness of the limitations and dangers of such automated text analysis systems and what such systems cannot do. Critiques in Ericsson and Haswell's (2006, pp. 5-6) collection provide the following reasons for this tension. Writing teachers and researchers are worried because they question

- whether such systems can be gamed or fooled, and whether students can break these systems
- whether machine analysis programs can fully understand the meaning of texts
- how students would react when they find out their work has been evaluated automatically
- how closely such software matches the careful evaluation of writing teachers
- where automated text analysis leads the teaching profession, would tutors have greater or less control over courses.

Scepticism and criticism have accompanied automated essay analysis over the years (Attali & Burstein, 2006). In 2013, the international journal of *Assessing Writing*, a referred

journal on the assessment of written language, dedicated a whole special issue to the automated assessment of writing. This special issue described this critique as follows:

> Those who have devoted their careers to the study, teaching, and assessment of writing view writing not as a simple behaviour easily evaluated through computer algorithms; rather, writing researchers view composition as a rich and nuanced activity informed by myriad discursive and non-discursive purposes, practices, and perspectives (Elliot & Williamson, 2013, p. 1).

"The past has been marked by an argument rather than dialogue: one side is the researchers and developers of such systems and the other side is the critique of such systems from writing teachers and researchers" (Shermis & Burstein, 2013, p. ix). Since neither side has yet to reach a compromise, this issue will be further discussed in the literature review, as this thesis investigates the use of automated analysis of meta-discourse in student writing to support tutors' essay assessment practices. Therefore, it is important to acknowledge this tension and unpack it further.

In addition to the technical aspect of this thesis, i.e. automated analysis of meta-discourse in student writing which has been introduced above, the following section introduces the learning aspect of this thesis.

1.5 Learning analytics

The growing number of data sets and the opportunity to study these using computational techniques has led to the development of *analytics*. The term analytics refers to the processes of studying such data sets and analysing them to measure, improve, and compare the performance of individuals, programmes, departments, or institutions (Norris, Baer, & Offerman, 2009). Analytics technology aids decision makers to find the best course of action by evaluating large data sets (Brown, 2011). "Analytics is the process of developing actionable insights" (Cooper, 2012, p. 3), and *action analytics* refers to "analytics capabilities and practices that are powerful, immediate, and lead to outcomes that are useful to a wide variety of stakeholders" (Norris *et al.*, 2009, p. 1).

Analytics is used in business and science to describe computational support for capturing digital data trails to provide rapid feedback, timely interventions and to help inform decision-making processes. *Learning analytics* brings this concept into an educational context and considers how learning data should be analysed to improve learning and the environments in which it occurs, based on the assumption that big data and analytics can add value to education by shaping its future (Siemens & Long, 2011).

The Society for Learning Analytics Research defines learning analytics as "the measurement, collection, analysis and reporting of data about learners and their contexts, for purposes of understanding and optimizing learning and the environments in which it occurs". Digital data trails produced by learners, such as log-in information, rates of participation in specific activities, the amount of time students spend interacting with online resources, etc. can be used to understand what happens during learning processes and can be useful to find out what kind of improvements should be carried out by educators (Siemens & Long, 2011). Additionally, analysis of learner-produced data can provide detailed information about the potential problems experienced by students who might need additional support (Siemens & Long, 2011). It can help not only learners by showing their own progress and learning habits back to them but also educators to understand the impact of changing various elements in learning processes.

The type of data gathered varies by institution and by application, but in general it includes information about the frequency with which students access online materials or the results of assessments from student exercises and activities conducted online. Since the focus of this thesis is on the automatic identification of discourse elements in students' writing, learning analytics based on discourse elements will be a main theme for this thesis, which is explained in the next section.

1.5.1 Discourse-centric learning analytics

Most learning analytics applications provide quantitative information about learners, e.g. based on how many times they have logged in to learning platforms, viewed a forum post, and replied to it. However, learning analytics can move beyond reporting these simple quantitative logs, and provide information on the quality of these contributions students made (Buckingham Shum, Knight, & Littleton, 2012). One interest for learning analytics is in its potential for the analysis of discourse data (Buckingham Shum & Ferguson, 2012).

> Researchers are beginning to draw on extensive prior work on how tutors mark essays and discussion posts, how spoken and written dialogue shape learning and how computers can recognize good argumentation, in order to design analytics that can assess the quality of text, with the ultimate goal of scaffolding the higher order thinking and writing that we seek to install in students (Buckingham Shum et al., 2012, p. 6).

Discourse-centric learning analytics is a term first defined by De Liddo, Buckingham Shum, Quinto, Bachler, and Cannavacciuolo (2011) in the first Learning Analytics and Knowledge conference (LAK). De Liddo *et al*. (2011, p. 6) motivated a learning analytics that focuses on "learners' discourse as a promising site to identify patterns of meaningful learning". Their work identifies the rhetorical attitude of learners towards discourse contributions, like arguments supported and rejected by learners, the evidence they used for such arguments, and emerging questions.

Following this, the first discourse-centric learning analytics (DCLA) workshop (Buckingham Shum *et al*., 2013) held at the third LAK conference proposed a mission statement for DCLA: "to devise and validate analytics that look beyond surface measures in order to quantify linguistic proxies for deeper learning" (Ferguson, De Liddo, Whitelock, De Laat, & Buckingham Shum, 2014, p. 1). In 2014 as part of the fourth LAK conference, the second DCLA workshop was held with a focus on the intersection of learning analytics research, theory and practice: "once researchers have developed and

validated discourse-centric analytics, how can these be successfully deployed at scale to support learning?" (Ferguson *et al*., 2014, p. 1).

Learning analytics with a focus on the use of discourse to support learning and teaching are being developed at the intersection of fields such as automated assessment, learning dynamics, deliberation platforms, and computational linguistics. Ferguson *et al*. (2014) questioned what steers such developments towards the category of learning analytics, as opposed to research that sits in any of the other categories: their use or potential to generate actionable intelligence specifically in the context of learning, such as helping educators to understand significant discourse patterns.

The definition for this addition to learning analytics came from Knight and Littleton (2015, p. 17): "DCLA focuses on analytics to support high quality discourse for learning contexts; it consists of analysis of discourse data, creation of effective feedback to learners and educators, and the validation and theorising of our analytic techniques".

The 'D' in DCLA stands for discourse coming not only from student writing but also from social interactions, online discussions, forum posts, and exploratory dialogue. As the DCLA workshops had already produced a couple of papers on extended student writing, at the sixth LAK conference a new workshop was held specifically focusing on discourse in student writing, called 'Critical perspectives on writing analytics' (Buckingham Shum, Knight, et al., 2016). "Broadly defined, writing analytics involves the measurement and analysis of written texts for the purpose of understanding writing processes and products, in their educational contexts" (Buckingham Shum, Knight, et al., 2016). This workshop therefore focused on analytics that can help to gain a better understanding both of the writing process as well as the final product, and of the pedagogical context in which writing analytics should take place, i.e. how to embed writing analytics meaningfully within a pedagogical context.

DCLA as a sub-area of learning analytics does not only take an interest in computational-analytic techniques for discourse but also in the explicit learning implications of those techniques (Knight & Littleton, 2015); this is why this thesis is part of the field of discourse-focused learning analytics. The next section sets out the thesis structure for the remaining chapters.

1.6 Structure of this thesis

The current chapter has provided an introduction to the aim, approaches and concerns of this thesis, as well as to three fields, and this thesis sits right at the intersection of those fields: academic writing, automated analysis writing and learning analytics. The rest of the thesis is organised as follows:

Chapters 2 & 3: Literature Review

The next two chapters locate this thesis more firmly within the existing literature, and review the previous theoretical and empirical work on which it is based. The literature review has been divided into two chapters: academic writing and automated assessment of writing. In particular, it investigates what really matters in student academic writing, what educators think good student writing is and how this differs within specific disciplines. It also provides detailed accounts of key approaches to automated text analysis.

Chapter 4: Methodology

The fourth chapter focuses on the research aim, philosophy and datasets used for analysis: their selection, collection and rationale. The mixed-methods approach that is adopted in this PhD research is discussed along with a justification for the choice of the methodologies. The research methods used in this thesis – together with regression analysis, focus group, and one-to-one interviews – are described and their advantages, disadvantages and ethical concerns are considered.

Chapter 5: The Automated Analysis of Student Writing

The first analysis chapter addresses the question ‘To what extent can the automated rhetorical parser XIP be used to identify indicators of good academic writing in undergraduate student essays from different disciplines, as judged by the essay grade?’. It explores how well automated meta-discourse analysis works across disciplines through quantitative regression analysis by using various datasets, student writing and their essay mark, from different disciplines and levels.

Chapter 6: One-to-one Interviews with Markers

The second analysis chapter asks ‘How do educators define the attributes of good student writing and to what degree can the automated rhetorical parser XIP identify the presence of these attributes?’ It investigates how educators define the quality of student writing, what they give credit for when marking a student essay, and to what extent automated rhetorical analysis can capture these.

Chapter 7: Focus Group

The third analysis chapter addresses the question, ‘How congruent is the XIP analysis of student essays with educators’ judgement of quality?’ It explores the possibilities of integrating an automated rhetorical analysis into educators’ essay assessment practices.

Chapter 8: Making use of the output

The final analysis chapter addresses the question**,** ‘In what ways should the output of the XIP analysis of student essays be delivered to educators so that they make use of this output to give feedback on student essays?’ It explores what educators would require to make use of the output of such automated analysis so that it would support their essay assessment processes. The user and system requirements are investigated through unstructured interviews where the participants are selected for their expertise in this area**.**

Chapter 9: Discussion and conclusion

The final chapter employs the findings reported in Chapters 5 to 8 in order to answer the main research question, 'To what extent can the automated rhetorical parser technology be used to identify indicators of good academic writing in undergraduate student essays and to support educators' feedback processes?' It considers how successful the research has been in fulfilling its aims, and goes on to identify its original contributions before outlining an agenda for future research.

2

LITERATURE REVIEW:
Academic Writing

2.1 Introduction

"Writing development is painstakingly slow
because academic writing is never
a student's mother tongue."

(Sommers & Saltz, 2004, p. 145)

This thesis focuses on 'student academic writing', which refers to argumentatively written student essays that require students to acknowledge the literature, the debate between researchers and then build on these ideas with a critical eye. The genre of writing is 'argument writing' which is the most common genre that undergraduates have to write (Mei, 2006) particularly in the arts, humanities and social sciences (Hewings, 2010). Therefore, 'student academic writing' in this thesis does not deal with the other genres of writing that students might engage in during their studies such as creative writing, interpretive writing, descriptive writing, reflective writing, case study reports, and factual technical reports. Clearly, targeting all kinds of undergraduate student writing would be complex and unsuitable considering the different requirements each of these writing types require. Therefore, this thesis focuses solely on argument writing as a genre.

2.2 Academic writing

In the academy, knowledge is built and produced primarily in written form (Bazerman, 1988). The written text, published in a journal or book, is a definitive form of claims and

arguments, which build on prior work, leading to future claims (Bazerman, 1988). When writing an academic piece, writers make sense of other writers' points of view and decide when to attribute an idea to another writer, which is appropriate to the academic community to which they belong (Hyland, 2002; Ivanič, 1998). Academic writing is an act of self-representation, personal conviction or 'voice' in the writing which not only conveys disciplinary content, but also carries a representation of the 'author identity' which is the author's stance towards the subject (Andrews, 2003; Hyland, 2002).

Journal paper writing, book writing or other sorts of experienced research writing is not the same genre as undergraduate writing, as they require different discursive moves. A discursive move as part of a written discourse contains a set of communicative acts related by a goal or theme. Discursive moves can serve authors to motivate, instruct, explain, compare, or summarise depending on the goals. For instance, in the field of biochemistry, authors can use discursive moves to provide an introduction to their study, stating the purpose, describing the procedures, presenting the findings, and stating the limitations (Kanoksilapatham, 2005). As an illustration, the following sentences are taken from an introduction of a biochemistry journal article, conveying the author's rhetorical moves as:

- making topic generalisation: "Protein export pathways are less well characterized, although..." (Kanoksilapatham, 2005, p. 274).
- indicating a gap: "Consequently, how related the serotonin N-acetyltransferase catalytic mechanism will be to that of other superfamily members is unclear" (Kanoksilapatham, 2005, p. 275).
- presenting findings: "Our results show that U2snRNP is functionally associated with the E complex and is also required for its assembly" (Kanoksilapatham, 2005, p. 276).

The way in which authors use discourse moves to convey similar meaning might vary in different disciplines. Additionally, the types of discourse moves used might vary between disciplines; in specific disciplines certain moves might not be widely preferred. For

example, in the field of economics authors might not prefer to indicate a gap in their introductions, whereas in medicine they might do so extensively (Ryvitytė, 2003). Although experienced research writing authors are generally expected to present new arguments, indicate gaps in their fields, and present their studies and findings, many of these discursive moves are not expected to be found in undergraduate student writing. The discourse moves introduced above are examples taken from journal papers, but discourse moves are also part of student writing. For instance, Ivanič put the phrase 'writer identity' at the centre of undergraduate writing (Ivanič, 1998; Lillis & Turner, 2001). The word *identity* here explains how student academic writing should be perceived. It needs to move beyond simply reporting findings or expressing others' ideas, to employing accepted rhetorical resources in a particular genre and social community (Hyland, 2002). This enables students to evaluate, and internalise what has been said in that social community in order to develop their own perspective, 'voice' within the community, allowing them to adopt more sophisticated skills (sophisticated writing skills are higher order writing skills such as argumentation or organising ideas coherently, whereas lower order writing skills are things like spelling, punctuation, or grammar).

"Writing encompasses a wide range of skills, from the mechanics of punctuation and spelling to the systematic or even creative development of ideas. The higher order skills of communication necessarily involve critical thinking and problem solving..." (White, 1993, p. 3). Students do not add anything from their identity when the issue is spelling and sentence structure; they do what they should specifically in order to produce grammatically correct sentences. When they develop arguments, conduct research, or solve problems, however, they think of themselves as individuals (White, 1993), and their identity plays the role. The significance of creating one's personal views and opinions in the making of text has been signalled several times through seminal works on academic writing, see for instance (Bakhtin, 1981; Clark & Ivanič, 1997; Fairclough, 1992).

2.2.1 Academic discourse community

While academic writing allows student writers to develop personal views, opinions and identity in their disciplinary area, it also enables them to adopt the practices and discourses of a community in which they are involved. Academic writing provides a way of expressing perspectives and ideas in ways similar to other members of this community (Hyland, 2002). As defined by Swales (1990), a discourse community has six defining characteristics:

1) a broadly agreed set of common public goals
2) mechanisms of intercommunication among its members
3) it uses its participatory mechanisms primarily to provide information and feedback
4) it utilises and hence possesses one or more genres in the communicative furtherance of its aims
5) in addition to owning genres, it has acquired some specific lexis ("the total stock of words in a language; the level of language consisting of vocabulary, as opposed to grammar or syntax" (Oxford English Dictionary, *lexis*))
6) it has a threshold level of members with a suitable degree of relevant content and discoursal expertise.

What it means to be a member of a specific academic discourse community, constituted by a range of values, assumptions and practices, is generally found challenging by most higher education students embarking on their studies, who need to learn the expectations and norms of academic discourse communities of their disciplines (Ivanič, 1998). Teaching students to write for their undergraduate courses initiates them into the academic discourse community (Bizzell *et al*., 1982). A discourse community is a group of people who generally have a shared interest, goals or purposes and use written communication to achieve these goals (Borg, 2003; Swales, 1990).

Students might not be a part of a discourse community that relates solely to argumentative writing; but to position themselves within an academic discourse community, newcomers need to learn the expectations of this community with regard to written texts, and this has implications for the teaching of writing within academic communities (Atkinson, 1997; Borg, 2003; Swales, 1990; Wenger, 1998). "There are different literacies in different contexts, so that students need to learn the specific characteristics of academic writing, and of the disciplinary culture into which they are entering" (Ivanič, 2004, p. 222). Consequently, educators should be aware of how academic conventions position students, and be sensitive to the struggles of novice writers (Hyland, 2002). That is why there is "an important consciousness raising task to ensure students understand the rhetorical options available to them and the effects of manipulating these options" (Hyland, 2002, p. 1111) so that such understanding will give an insight to the learners about how to control their writing and how to meet the challenges of participating in academic genres. To ensure that students can understand this, various research studies have been conducted, which are outlined in the following sections. These studies range from finding generic conventions for academic writing in general, to defining genre-specific expectations, needs and requirements of different academic communities. Before moving on to differences and similarities between disciplinary discourses, the next section will give an overview of the role of academic writing in higher education, and discuss what really is important in student essays and what the academy requires students to write.

2.3 Academic writing in higher education

While many undergraduates view academic writing as the "acquisition of correct information and right answers", it actually requires "analytical or argumentative thinking" (Bean, 2011). It is a process of joining conversations with people who are disagreeing with each other or who have shared interests in seeking answers to questions (Bean, 2011).

Therefore, academic writing in higher education enables professional advancement for university students as it nurtures thinking and reflection.

However, students' writing background dates back to school writing, which differs from student academic writing in higher education. "After their long experience with the schoolroom essay, and long contact with rules and pronouncements about good writing, university students suddenly face many examples of expression that contradict the schoolroom tradition" (Giltrow, Gooding, Burgoyne, & Sawatsky, 2014, p. 9). A significant difference between secondary school and university writing is that the university is a research institution that produces new knowledge through observation, experiments and interpretation of the existing world, "…the kind of writing that suits the schoolroom tends not to suit the university classroom" (Giltrow *et al*., 2014, p. 9). Although undergraduate students are learners at the university rather than researchers, they are taught by educators who mostly trained as researchers and who read and write research publications. Therefore, undergraduates are expected to adopt higher order writing skills such as argument writing and criticality, which are not taught or necessarily practised in secondary-school years.

Undergraduate students typically specialise in well-established disciplinary areas to become for instance a 'mathematician' or an 'historian'. In England, degree courses usually offer single honours degree to students, specialising through a limited number of courses. Therefore the role of writing in higher education has been limited to subject-specific courses that require this skill and it has not been considered necessary in disciplines such as mathematics (Lea & Street, 1998; Lillis & Turner, 2001). However, writing proficiency in general is an essential skill that higher education students should learn for effective communication; and they should "[…] experiment with different kinds of writing because experimentation forces one to develop new forms of perception and thought, a new and more complex sensibility" (Sommers & Saltz, 2004, p. 128).

A common reason for the lack of academic writing modules in higher education is a belief that students already knew how to write before coming to the university (Ganobcsik-Williams, 2006). In England for instance, many students need a good pass in English language which covers essay writing in order to start a course in most of the traditional universities. Additionally, students who take advanced level (A-Level) courses between the age of 16 and 18 to qualify for university entrance have prepared for their degrees with discipline-specific A-Level courses. A student who wants to study law at university can take A-Level courses such as history, English language and German, including the writing requirements of these subjects. However, undergraduates, especially in their first years, find the university writing difficult which is an activity that takes more time than they expected (Sommers & Saltz, 2004).

No matter how many essays they wrote during their studies before coming to the university, it is essential to bear in mind that school and university essays are different genres, and the latter requires particular advanced skills. Moving from a school culture to an academic culture, when senior students are asked to look back and reflect on their first year, they might define the writing activity as *building a house without any tools* (Sommers & Saltz, 2004, p. 131). Therefore, no matter from what school culture they come, every student in higher education needs to acquire the academic writing skills appropriate to their level and discipline.

Lea (1995) suggests that in order to understand more about student writing, it is necessary to consider writing within an institutional or disciplinary context where students actually produce written text, rather than examining frameworks that conceptualise writing in terms of study skills. She also suggests that there is a gap between the expectations of academic staff and student in respect of writing, which needs to be considered when researching academic writing in higher education. Although Lea identified this gap in 1995,

researchers have also identified it in much more recent studies, see for instance (Giltrow *et al*., 2014; Walvoord, 2014).

The following sections therefore deal with studies of this gap with the aim of gaining a full understanding of what academics really value in student writing, what they care about most, what they expect to see, and how they actually assess undergraduate essays.

2.4 What really counts in student academic writing?

Learning in higher education involves new ways of understanding, interpreting and organising knowledge. Student academic writing is therefore more than punctuation and grammar. Academic literacy practices, which mean the ability to read and write within disciplines (Lea & Street, 2006), are key for university students to learn new subjects and to develop their knowledge about a new area of study.

Almost two decades ago, many academic staff claimed that standards of student literacy were decreasing each year, and that undergraduate students could no longer write (Lea & Street, 1998). Although 17 years have passed since Lea & Street (1998) made this assertion, academics are still upset about their students' writing abilities and they often say their students do not write well and that they really want them to write better (Giltrow *et al*., 2014; Walvoord, 2014). Even though several research studies have been carried out in the past two decades, this problem has remained. It is important to point out this unsolved problem and look for its underlying reasons. One reason for the problem is given as being the contrasting views between students and academics regarding the expectations and interpretations of written assignments (Norton, 1990; M. Lea & B. V. Street, 1998; Lillis & Turner, 2001; Andrews, 2010). The following sections provide details about studies designed to understand the views of both students and academic staff regarding academic literacy practices.

2.4.1 Mismatched expectations: students vs. tutors

The distinction between school culture and research culture in higher education discussed above explains why academics complain about their students' writing abilities (Wingate, 2012). There should be a smooth transition from school culture to research culture that enables student writers to '*join conversations'* (Bean, 2011; Giltrow *et al*., 2014) with the academic community they belong to, whose researchers disagree with each other and/or have a shared interest in finding out answers to questions. When they learn to be part of the conversation by acknowledging others' opinions to develop their own perspective, student academic writing is more likely to satisfy the tutors. Although most students enter university without this ability, they can gain this skill and become part of the research culture if they are nurtured well (Giltrow *et al*., 2014). For instance, the four-year longitudinal study that Sommers and Saltz (2004) carried out with 422 students from the entire Harvard Class of 2001 covering all disciplines and subject areas contributed to the knowledge that students can gain this skill during their studies as they make their way to their final year.

Lillis (2002) shares how, as a tutor, she witnessed the struggles of students whose difficulties primarily centred on academic writing. She summarises the problem as students trying to '*write within rules of the game*' without knowing what the rules actually are. There are various studies showing how these rules differ in the perceptions of tutors and students. Although the Sommers and Saltz (2004) study provides students' perceptions from the Harvard class of 2001, there are some studies that investigate both sides with the aim of identifying the differences between the views of tutors and students on what university writing is. Lea and Street (1998), for instance, carried out a study within one new and one traditional university in England that involved semi-structured interviews with academics and students, participant observation of students' study group sessions,

linguistically-based analysis of students' writing samples, the feedback given to student essays, and hand-outs/guidelines provided to students on essay writing. This showed clearly that students and their tutors have different expectations and, more importantly, perceptions of how student academic writing should be realised. A similar but smaller-scale study was carried out by L. S. Norton (1990) that also shows there is a clear mismatch between students' and tutors' perceptions about the most important criteria for essay assignments. She found that students are more concerned with the *content* (factual information) whereas tutors are more concerned with the *argument* (perspective/viewpoint/stance towards the topic). Similar studies have been conducted on contrasting conceptions of essay writing between students and their tutors (see for instance Hounsell, 1997; McCune, 2004; Levin & Wagner, 2006; Lavelle & Bushrow, 2007; Fairclough, 2014).

Tutors' views

L. S. Norton (1990) interviewed six tutors with lecturing experience ranging from one year to twenty years and who marked students' essays. She wanted to understand how they marked these essays, and what they were specifically looking for. New lecturers seemed to have high expectations about student capabilities. Lecturers agreed that when they get more experienced, they become more generous and look for positive things rather than negative things in an essay. When asked what they actually looked for while marking these students' essays, they listed the key expectations and criteria as follows:

- whether students have actually understood what the essay question is about and whether they have actually answered it
- whether the essay is referenced properly
- whether students include evidence of wider reading rather than just the lecture notes

- whether the essay contains constructed arguments which are clearly set out by the student.

Tutors had a general concern regarding the way that essays are constructed, how an argument is formulated, and whether the essay is relevant to the topic. Some answers also showed that there are obvious individual differences in marking, as one of the tutors mentioned the ‘content’ as a major criterion whereas the other tutors did not view it as such. However, Norton’s sample was limited.

Her limited initial study with a sample size of six was later built upon by others. For instance, Lea and Street (1998) carried out research at two universities to investigate the misconceptions between tutors and students. They conducted ten interviews with staff in one university and 21 students. At the other university, 13 members of academic staff and 26 students were interviewed in the same way. Interviews with the academic staff in Lea and Street's (1998) work suggest that tutors actually have fairly defined views regarding the elements of good student writing. These range from generic attention to syntax, punctuation, and layout, to structure, argument and clarity. The interview results showed that the disciplinary background has a clear impact on what academic staff think that the most important elements of student writing are. For instance, in the subject area of history, tutors think the use of evidence is important, whereas in English clarity of expression is the most important element. This difference can also lead to problems especially when academic staff teach in programmes which integrate various disciplinary approaches. Researchers noted that most of the interviewed staff were mainly influenced by their own subject disciplines when assessing the student essays in any subject area.

As a key element of student writing, most of the academic staff interviewed mentioned ‘structure’ and ‘argument’ in Lea and Street's (1998) study. Staff had a common belief that these are the crucial elements of a successful piece of writing. There was less certainty, however, when it came to describing what makes a well-argued/structured, written piece.

Tutors were able to identify a successful essay, but could not describe why another lacked structure or what a well-developed argument looks like in a written assignment. One member of staff mentioned in Lea and Street's (1998) work that a good student essay should have an introduction that sets the scene and a main body that covers issues highlighted in the introduction where students should be critical, evaluate, reach some synthesis and summarise. However, when this member of staff was asked to clarify what s/he meant by these terms, s/he could not explain further the terms 'critically analyse', 'evaluate' and 'synthesise'. Similarly, another member of staff shared that she knows a good essay when she sees it, but cannot describe how to write it. As a result of these interviews, researchers concluded that elements of successful student writing are not related to a set of generic writing skills, as academic staff think from their own academic world-view and knowledge. They give feedback on their students' writing based mainly on descriptive categories such as 'structure and argument', 'clarity' and 'analysis'.

Students' views

In order to investigate the mismatch between tutors' and students' expectations about essay writing, these researchers also worked with students from different disciplines. L. S. Norton (1990) surveyed psychology students two weeks after their essay had been submitted in order to investigate their feelings about essay writing, the strategies they used, their work routines including number of hours spent on preparation, the number and types of sources they used, the grades they expected to get and most importantly the criteria they thought tutors use when marking their essay.

A questionnaire asked students to rank in order the six most important criteria that they thought tutors would use when marking their essays. The following table (Table 2.1) shows the mean ranking of what students thought were the most important criteria for essay grading, which are compared with their tutors' actual criteria (in the table, number 1

represents the most important criterion; so the bigger the mean ranking, the less important the criterion).

Table 2.1 Mean ranking of students' and tutors' criteria for assessment of essays

Source: (Norton, 1990, p.423 –Table 14)

Criteria	Mean ranking	
	Students (N = 98)	Tutors (N = 6)
Answer the question	1.5	1.0
Content/knowledge	1.8	–
Relevant information	2.3	2.0
Understanding	2.4	1.3
Wide reading	2.9	3.7
Structure/organisation	3.3	2.8
Argument	3.5	1.5
Evaluation/own views	3.9	3.5
Presentation/style	4.2	3.5
English/spelling	4.8	4.0

Both tutors and students agreed on *'answering the question'* as the most important criterion. However, this was followed by 'argument and understanding' for tutors, whereas students thought 'content/knowledge' was the second most important criterion. This was not even ranked by the tutors. Most students thought the main purpose of the essay was content/knowledge structure (factual descriptive information). Only a quarter of them were concerned with the argument. This result replicates Hounsell's (1984) study in which history students defined essay writing as a viewpoint and arrangement, whereas their tutors considered an essay to be an argument.

When students were interviewed, the common view was that they did not support the notion of generic and transferrable writing skills across the university. Students agreed that the most difficult thing is switching from one subject to another, and knowing how they are meant to write in each subject, and also for each tutor, as '*everybody seems to want something different'* (Lea & Street, 1998). Moreover, even in the same course module, different tutors had different opinions about the conventions and various comments on the appropriateness of the written piece. Like academic staff, students know that presenting an

argument and structure is important. However, they have problems in understanding when they had achieved this successfully in their writing. Most of the interviewed students had a common complaint about the negative feedback and low grades for essays that they believed were well constructed and appropriate to the subject area. Hounsell's (1997) interviews with psychology and history students regarding contrasting conceptions of essay writing yielded similar results.

Lea and Street (1998) also shared an example to show how feedback varies from one subject area to another. One student wrote two essays within the same week with the same style and manner for two different courses, history and anthropology, and received completely different feedback regarding the argument and structure. For history, which was his major degree, he received an excellent grade and feedback saying this was a '*…carefully argued and relevant essay'*. For the anthropology essay, he received a very low grade and received negative feedback claiming the student had a problem with the essay which did not have any argument and structure. The tutor even strongly suggested that he should visit the university study centre and make an enquiry about essay-writing clinics. The researchers interpreted this to mean that what makes an argument for one tutor might not be the same for another, and the anthropology tutor did not notice the linguistic and structural devices used by the student to indicate an argument.

More recently, Giltrow *et al*. (2014), emphasised the significance of carefully comparing writing with its own discipline as well as with other disciplines; as this is particularly important for guiding students properly.

> (When a student writes like an historian in philosophy class, the professor will notice). Genre theory tells us that the most of what people know about their own ways of writing and speaking is tacit, that is unspoken and undeclared. […] this tacit knowledge can be very awkward in its expression, even misleading, and also tending to unhelpful generalities: *make sure your essay has a coherent argument*. (Giltrow *et al*., 2014, p. xiv).

There are clear inconsistencies in the ways that tutors from different disciplines assess their students' essays. In order to gain a better understanding of tutors' assessment regimes, the

next section provides an overview of the quantitative aspect: what do essay grades actually tell us about good student academic writing?

2.4.2 What do essay grades tell us?

As Lillis and Turner (2001) said, students' written academic texts should be central for anyone who aims to explore what is involved in student academic writing. Therefore, this section covers the quantitative research that focuses on what the grades of written student texts tell us about their quality.

In Norton's (1990) study, a content analysis of twenty student essays was carried out: ten of the highest scoring and ten of the lowest scoring. This study showed what grades can tell about the requirements for a well written essay. High-scoring essays contained more research-based information and significantly less factual descriptive information than low-scoring essays. The number of references appeared to have a major effect on the final grade as the more references cited, the more research-based the essay appeared to be, and the higher the final mark. Factual descriptive information was significantly higher in low scoring essays. The findings of this study may be consistent enough to suggest these claims, but these findings are limited to first year psychology students at one university in 1990. It is therefore essential to note here that these claims require further investigation as the analysis of twenty essays from one discipline at one university does not represent a generalisable dataset, the results of which are necessarily not generalisable.

Argument has to go further than just presenting one's own view (Elander *et al*. 2006). Branthwaite, Trueman, and Hartley (1980) found that students were much more likely than lecturers to emphasise the need for originality in their opinions in essays, and students who believed that presenting their own opinions is the most important factor obtained lower grades for their essays than those who did not. These studies showed clear differences

between students' perceptions of how their educators assess their studies and gaps regarding the core assessment criteria of student academic writing.

With regard to the grades, it is essential to look at the generic assessment criteria for undergraduate writing, because the features of good student writing can be identified through talking to academic tutors who mark these essays, looking at their assessment criteria (rubric) and essay grades. Assessment criteria are a set of standards that students have to adhere to get a certain grade (Norton, Pitt, Elander, & Reddy, 2009). The assessment criteria for any given assignment are a list of specific aspects that tutors look for when marking students' work. Although assessment criteria can vary in different disciplines, Norton *et al*. (2009) showed that most tutors agree that the basic core assessment criteria for essay writing include the following:

- answering the question
- structuring the essay
- demonstrating understanding
- developing an argument
- using evidence
- evaluating sources
- use of written language.

Before moving on to the differences between disciplinary discourses, the next section will give a brief overview of arguments and studies on the conventions of academic writing. If the assessment criteria carry similar points for various genres and if there are some studies claiming similarities between high-scoring essays and low-scoring essays, can there be standards for good student academic writing? The aim is to clarify the current consensus on what the attributes of good academic writing are, the debate as to whether or not they are discipline specific or applicable to academic writing in most or all disciplines, and related issues concerning how academic writing can best be taught.

2.5 Conventions of academic writing

Research into academic writing draws attention to the question of whether there are conventions for academic writing, or whether it is better to focus on discipline-specific discourses (Thonney, 2011). This section covers widely accepted conventions (through the literature from English-speaking countries), how these conventions can best be taught, and indeed, whether they can be taught. The latter debate will be covered in the following section.

Some academics believe that it is impossible to define a set of generic skills which can be learned and applied to all writing activities and genres, because writing conventions vary across academic disciplines and discourse communities (Russell, 1995; Wardle, 2009). Others, including some linguistic scholars, have shown that patterns prevail across diverse forms of academic writing (Bazerman, 1988). Although there are some variations between disciplines in terms of the ways in which scholars inquire, prove and select research methods, it has been argued that some principles are consistent, and appear in all sort of academic writing forms regardless of the discipline (Bennett, 2009; Thonney, 2011).

In order to determine which rhetorical features appear and in what form they are represented in research articles written by experienced academics, Thonney (2011) analysed 24 randomly selected peer-reviewed journal papers – four articles from each of six different disciplines: psychology, sports medicine, biology, marketing, literature, and engineering. Thonney's (2011) selection of both hard and soft disciplines (hard and soft disciplinary groupings are explained in the next section) creates a wide ranging dataset for relatively generalisable results. Although journal paper writing requires different discursive moves and is a different genre to undergraduate writing, it is useful to look at what experienced writers do in their academic papers so that what is expected from undergraduates who moved from the school culture to the academic culture can be

interpreted more effectively. Her analysis found six standard moves in academic writing, which appeared in every article sample (Thonney, 2011, p. 348):

1. *Writers respond to what others have said about their topic*. Rather than just showing they have done the reading by stating other authors' work, writers analyse the arguments and claims of other writers, synthesise what has been said and establish arguments based thereon (this is what many students struggle with).
2. *Writers state the value of their work and announce the plan for their papers.* Scholars draw attention to unresolved issues in their own or others' work in order to motivate editors and reviewers (or professors, in the case of student writers) to read the current paper.
3. *Writers acknowledge that others might disagree* with the position they have taken.
4. *Writers also adopt a voice of authority*.
5. *Writers use academic and discipline-specific vocabulary*.
6. *Writers emphasise the evidence, often in tables, graphs and images*.

Although there are some variations amongst writers, there is also useful general knowledge, and several techniques and conventions of academic writing that can be followed by writers and especially novice writers (Thonney, 2011). Students can adapt these into their writing. There are studies showing disciplinary variation in academic writing, but the importance of understanding the conventions that are applicable to various disciplines should not be neglected. There is an increasing need for studies that tell what academic writing has in common, what the common structure is and the must-have patterns are; so that novice writers can adapt such techniques (Thonney, 2011).

On the question of teaching, some scholars argue that academic conventions cannot be taught as no one can learn to write academically by following a definitive rule-set (Berkenkotter & Huckin, 1993; Freedman, 1993; Lynch-Biniek, 2009). On the other hand,

others report the benefits of teaching such academic writing patterns and conventions to students; in order to help them to understand the mystery of writing (Birkenstein & Graff, 2008; Williams & Colomb, 1993). Once students have an overview of writing conventions, they can adapt these into their area. There are some broad definitions of good student academic writing that have been given by some researchers as well as rubrics, see (Walvoord, 2014). Academic conventions can be taught; however, it would not be wise to ignore the fact that there are clear differences between some disciplines and good student writing in biology might look totally different from writing in philosophy. Therefore, in order to gain a better understanding of disciplinary differences in academic writing, the next section discusses the discipline-specific features of good student academic writing.

2.6 Disciplinary variation in academic writing

"[…] the styles of the different disciplines share many features, but the differences are meaningful and have consequences" (Giltrow *et al*., 2014, p. 10). While some researchers have focused on conventions in academic writing; mainstream research into this form of writing is mainly concerned with how writing conventions vary between subject specific academic communities, what the discipline-specific conventions could be and how writers need to adapt themselves to the rhetorical options and genre of that community.

For many decades, there has been considerable research conducted to improve writers', specifically student writers', learning strategies, and outcomes can be discipline dependent. Discipline-based studies have been conducted since the mid-20th century. One of the leading scholars of this area, Biglan, argued that there are three dimensions of academic subject matter that are perceived by scholars (Biglan, 1973a). The first dimension differentiates hard sciences, engineering and related subjects from social sciences, education and humanities. He labelled the dimension as *hard-soft* where *soft* stands for social sciences, humanities etc. and *hard* for physics, chemistry, engineering, and so forth.

His second dimension is called *pure-applied* in which the degree of concern is the application of disciplinary knowledge. The third dimension is *life or non-life* which is distinguished by whether the discipline is concerned with living organisms or not. As an illustration, Table 2.2 gives examples of disciplinary groupings based on Biglan's three dimensions (Biglan, 1973b). Accepting Biglan's disciplinary groupings, this thesis classifies the disciplines as hard and soft in the rest of the thesis.

Table 2.2 Disciplinary Groupings based on Biglan's Dimensions

Source: (Biglan, 1973b, p.207 –Table 2)

Task area	Hard		Soft	
	Nonlife system	Life system	Nonlife system	Life system
Pure	Astronomy Chemistry Geology Math Physics	Botany Entomology Microbiology Physiology Zoology	English German History Philosophy Russian Communications	Anthropology Political science Psychology Sociology
Applied	Ceramic engineering Civil engineering Computer science Mechanical engineering	Agronomy Dairy science Horticulture Agricultural economics	Accounting Finance Economics	Educational administration and supervision Secondary and continuing education Special education Vocational and technical education

Similarly Kolb studied disciplinary differences among academic communities (Kolb, 1981). He built his stance on Biglan's approach and studied disciplinary differences from the perspective of individuals' cognitive learning styles. Based on this approach, he created a four-dimensional experiential learning model which is shown in Figure 2.1. He described four stages in this model: 'Concrete experience', meaning learners need to "involve themselves fully, openly and without any bias in new experiences"; 'reflective observation' which means they need to "observe and reflect on these experiences" from various perspectives; 'abstract conceptualization' which means learners need "to create concepts that integrate their observations into logically sound theories"; and an 'active experimentation' stage where learners need to use these theories for decision making and problem solving (Kolb, 1981).

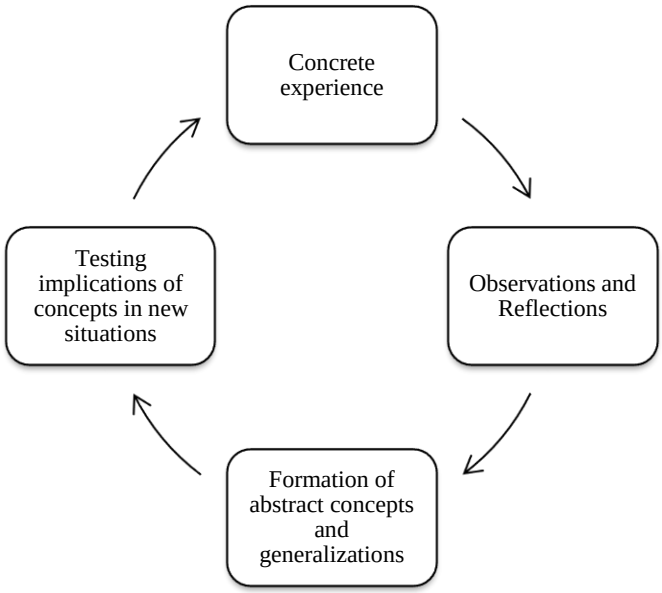


Figure 2.1 The experiential learning model

Source: Kolb, 1981, p.235 – Figure 1

Kolb summarised this model into two polar opposite dimensions: *concrete-abstract* and *active-reflective*. Based on Biglan's theory, Kolb concluded that these two dimensions are differentiated sharply among academic disciplines (Kolb, 1981).

Becher explored some of the key differences between disciplines, and the implications of such distinctions for education (Becher, 1994). Becher studied research norms and practices in two studies which were based on 350 semi-structured interviews with academics and research students. Becher combined Biglan (1973b) disciplinary groupings and Kolb (1981) experiential learning cycle and sought to classify disciplinary areas by grouping (Table 2.3).

Table 2.3 Becher's broad disciplinary groupings

Source: (Becher, 1994, p. 152- Table 1)

Biglan	**Kolb**	**Disciplinary areas**
Hard pure	Abstract reflective	Natural Sciences
Soft pure	Concrete reflective	Humanities and social sciences

Hard applied	Abstract active	Science-based professions
Soft applied	Concrete active	Social professions

A significant body of research into disciplinary variations builds on the work of these three scholars, significant works in this area include: (Hyland, 2004; Nesi & Gardner, 2006; Neumann, 2001; Neumann, Parry, & Becher, 2002; North, 2005; Parry, 1998).

Given the evidence that disciplines have such different expectations and norms, one might expect these to influence the associated writing of academics, and hence, students. "Writers in different disciplines represent themselves, their work and their readers in different ways, with those in humanities and social sciences taking far more explicitly involved and personal positions than those in the sciences and engineering." (Hyland & Bondi, 2006, p. 36). Hyland and Bondi (2006) described the differences between soft disciplines on the one hand, and hard science disciplines on the other, based on the assumption that hard and soft disciplines differ quite clearly along many dimensions of variation. One of the dimensions of variation is the use of *citations*, the way that writers construct their arguments and claims based on other writers' work and frequency of those citations. Citation is taken into account to show the difference in the ways in which writers' knowledge-level claims are constructed: whether or not they are based on the evidence presented in the paper, or on the pre-established arguments put forward in others' studies. One study showed that, in the humanities and social sciences, literature is open to greater interpretation, drawing on the work of others to elaborate on the context via citation. This study was based on 80 research articles from eight disciplines. It found that articles in "philosophy, sociology, marketing and applied linguistics together comprised two thirds of all the citations in the corpus, twice as many as science disciplines, with engineering physics well below the average" (Hyland & Bondi, 2006).

Discipline Citations	*Av. per paper*	*per 1000 words*
Sociology	104.0	12.5
Marketing	94.9	10.1
Philosophy	85.2	10.8
Biology	82.7	15.5
Applied Linguistics	75.3	10.8
Electronic Engineering	42.8	8.4
Mechanical Engineering	27.5	7.3
Physics	24.8	7.4
Totals	67.1	10.7

Figure 2.2 Number of citations by discipline within 80 research articles

Source: Hyland & Bondi. 2006, p. 26 - Figure 2

What Hyland showed was that even the academic conventions defined by Thonney (2011) have discipline-specific aspects to them. Thonney (2011) argued that "writers respond to what other academics have already argued about their topic", but Hyland showed that this is not as common in the hard sciences as in the soft disciplines, although this does not change the fact that both disciplines have similar conventions, only the level of use is different.

Within a similar context, North (2005) conducted a three-year longitudinal study to develop an account of how students' writing is shaped by disciplinary background. Her study was specifically on student essays; which makes it particularly relevant in the context of this thesis. North's study was conducted with students from soft and hard science backgrounds (using the term 'soft' and 'hard' as defined by Biglan above), who took the same undergraduate course – the history of science – , requiring essay writing as an assignment. North (2005) found that the average essay mark for arts students was significantly higher than that awarded to science students. Linguistic analysis of the essays showed that there was a significant difference between what she termed "theme/rheme" structures. When using this structure, the first section of the sentence (theme) provides an interpretation for the subsequent clause (rheme). Done well, this helps the reader grasp the author's message, what s/he is going to say and his/her critical stance. North (2005)

reported that, compared to science students, arts students used orienting themes more in their essays which commented on the following proposition (rheme). She proposed that in the soft disciplines, writers are far more likely to construct interpretations based on others' arguments to establish their knowledge, whereas writers in the hard disciplines construct knowledge based on numeric evidence: the claim is that 'the facts speak for themselves' and require little elaboration. In short, North's (2005) claim is that arts-oriented knowledge demands a more rhetorical performance, in contrast to the more straightforward representation of reality assumed by the harder disciplines.

Instead of directly accepting the statement that when judged by one specific set of single criteria, science students' writing does not get such high marks as art students', this thesis takes the position that writing up experiments is relatively straightforward for undergraduate students, as this writing structure has been taught in all science lessons since primary school. However, this point requires further investigation to gain a better understanding of what works well and what does not in different disciplines.

There is a need for interpretation of how argumentative writing in different disciplines occurs. Argumentation at undergraduate level in particular continues to be under-researched in the UK (Andrews, 2009). "While there has been a slow but steady rise in interest in writing across the disciplines, argumentation research at this level remains small-scale" (Andrews, 2009, p. 1). Considering the key requirement that tutors look for when assessing a student essay is defined as "argumentation" (Lea & Street, 1998), argumentation research is needed for student writing across the disciplines.

One of the studies has come from Andrews (2010), who discussed how argumentation occurs in different disciplines from history to science. His position was that, regardless of the discipline, all academic argumentation involves negotiating a new position or defending an existing one in relation to others. Andrews (2010) acknowledged the disciplinary constraints that play a major role in shaping the nature of an argument. Although historians see argumentation as central to the discipline and consider that "it

[argumentation] is the discipline" (Andrews, 2010), biologists do not see argumentation as key. What historians mean by argumentation is the process of reading sceptically, seeking evidence by putting together different aspects, questioning what happened and why, and conducting an argument at the end (Andrews, 2010). On the other hand, some biology lecturers' view is that students cannot know enough to argue a position during their undergraduate years when they are building their knowledge of the field. In his book, Andrews not only discussed biology and history but also discussed how argumentation is perceived in student writing in various other disciplines such as mathematics, electronics, art, humanities, social sciences, etc. through interviews with tutors from these disciplines. Such differences show that there is a need to consider and pay attention to disciplinary differences when studying and researching undergraduate student writing.

2.7 Chapter Summary

The first chapter of the literature review on academic writing has covered the following key points:

- Academic writing is an important skill that undergraduates should acquire.
- When they move from a school culture to an academic culture, students find undergraduate writing difficult.
- Students also struggle to understand the norms of university writing and cannot easily grasp what is expected from them.
- Argumentation is a key requirement for student writing.
- Argument writing requires more than fact telling. Students need to acknowledge others' points of views in their discipline and then establish their perspective within their academic discourse community.
- There is a mismatch between the expectations of tutors and students with regard to the written assignment.

- Like students, academic tutors experience problems in providing effective examples and feedback; therefore, they also require support.
- Academic writing includes similar structural elements in all disciplines, but these expectations are also influenced by subject area.

Student writing is central to assessment in higher education (Andrews, 2009). Argument writing as a genre lends itself to persuasive discourse: "ideas are paraded, supported by evidence, linked into meaningful sequences and commented upon in order to persuade the reader of the strength of the writer's position" (Andrews, 2009, p. 3). These qualities make it assessable, as it allows academic tutors not only to gauge the student's understanding, but also to differentiate between students; therefore it "is the genre par excellence for assessment in the academy" (Andrews, 2009, p. 3). Since it is central to assessment in higher education, the next chapter reviews the literature on the automated assessment of writing, which could be used to overcome the problems that tutors and students experience in this area, as explained above.

3

LITERATURE REVIEW:
Automated Assessment of Writing

3.1 Introduction

The best way to improve one's [academic] writing skills is to write, receive feedback from an instructor, revise based on the feedback, and then repeat the whole process as often as possible" (Burstein, Chodorow, & Leacock, 2003, p. 3). This cycle requires tutors to read and provide feedback on student essays, which can create an enormous workload (Burstein *et al.*, 2003). This problem led researchers to study ways of developing applications that can automatically analyse and evaluate essays for assessment purposes. Automated essay evaluation (AEE) is the process of evaluating and scoring written prose via computer programs (Shermis & Burstein, 2003).

Since writing is an activity that is deeply human, its association with computational formulations is double-edged (Elliot & Williamson, 2013). When algorithms are used by computers as the basis for assessing student writing beyond fluency or knowledge of grammar (Attali & Powers, 2008), there is a suspicion that technology can corrupt the essence of a fundamental human activity (Ericsson & Haswell, 2006; Herrington & Moran, 2012).

This chapter gives an overview of the history of AEE, some of the controversies that are salient to this thesis, how well such systems work, what current research directions are, and where there are gaps in the literature.

3.2 Early history

Research in automated essay assessment began in the early 1960s. One of the first applications, and the idea of using computers to grade essays, came from a former high school English teacher, Page (1966, p. 238), whose aim was to "rescue the conscientious English teacher from his backbreaking burden". His article was greeted with high scepticism as the idea of 'grading essays by computer' seemed nonsense and impossible for many people (Shermis & Burstein, 2013). Page's admonition was that undergraduates, especially those with poor writing abilities, have to write more to be better writers. This was associated with the assumption that their academic tutors were spending too much time marking papers and were burdened with hours of grading written assignments. Page argued this is the biggest impediment to more writing as overburdened tutors would ask limited number of drafts from their students. Therefore, he worked on persuading educators that there is a need for computers to grade essays automatically, and that this is feasible and promises to improve the quality of teaching (Page, 1966) . However, unlike sceptics, people who knew about disciplines such as linguistics and who understood about computers, considered his approach as a "delightful toy" at that time (Page, 1966, p. 238). Page's aim was to develop a computer program that could evaluate student writing "as reliably as human readers", saving time and providing timely, speedy, reliable feedback to students (Page & Paulus, 1968). However, access to computers was rare at that time apart from entering data via IBM's punch cards which was costly and not well-suited for the average person. What Page was suggesting sparked a controversial debate on the idea of 'replacing human markers with machines', but there was no easily accessible hardware or software that was capable of doing half of what Page was suggesting.

3.3 Automated essay scoring and evaluation applications

When Page and Paulus (1968) published *The Analysis of Essays by Computer*, they visualised a promising future of automated programs that could evaluate both the aesthetic

traits of essays and their content. However, it was not until the 1980s that such technology was accessible. As basic word processing systems became available in the Eighties, these systems made it possible to input text into automated essay assessment systems. By the early 1990s, the advances of the Internet and word processing systems made Page's idea of automatically evaluating student writing possible. With important developments in the following decades, several studies have been conducted on automatic essay scoring (AES) and automated writing evaluation (AWE); see (Foltz, Kintsch, & Landauer, 1998; Landauer, 2003; Shermis & Burstein, 2003) for detailed descriptions of AES and AWE. One of the widely known applications of AES is *E-rater*™. E-rater has been used by the Educational Testing Service (ETS®) for AES in the United States of America and became the first large-scale assessment tool to incorporate automated essay scoring (Attali & Burstein, 2006). It is a web-based system that provides automated scoring and evaluation of student essays. ETS detects errors in grammar, usage, and mechanics, identifies discourse elements in an essay, and recognizes elements of undesirable style.

Page also later worked on his innovative idea with Petersen in 1995, *The Computer Moves into Essay Grading: Updating the Ancient Test,* through ETS (Page & Petersen, 1995). Page and Petersen's (1995) collaboration with ETS allowed them to show the possibility of Page's initial idea: "in a blind test a computer can simulate the judgement of a group of human judges on a brand-new set of essays" (p. 561). However, Page and Petersen's (1995) argument that computer programs can do better jobs than human readers and that therefore one day computers may replace them at essay grading has lead an unresolved tension, which is discussed in the next section.

ETS is not the only testing service that uses automatic essay scoring systems, they have been adopted as a second or check scorer in widely known American exams such as the Graduate Record Examination (GRE®), the Test of English as a Foreign Language (TOEFL®), and the Graduate Management Admissions Test (GMAT®). AES systems are

also used as the primary essay scoring engine in various assessment and instructional products, including Accuplacer®, the Criterion® Online Writing Evaluation Service, Intellimetric®, and the Pearson Test of English™, which are examples of how AES has been developed towards the idea of Page, who supplied the spark that ignited the controversial arguments around using computers to grade essays.

3.4 Controversial views on AES: does it really work?

"… it has come to this.
The essay, the great literary art form that
Montaigne conceived and Virginia Woolf carried on …
has sunk to a state where someone thinks it is a bright idea
to ask a computer if an essay is any good." (Scott, 1999)

Automated essay evaluation, especially automated essay *scoring*, has been subject to significant controversy. On the one hand there is significant support for AES as "automated essay scoring and evaluation becomes more widely accepted as an educational supplement for both assessment and classroom instruction" (preface in Shermis & Burstein, 2003). There are several studies showing that AES systems work well, and studies reporting high agreement rates between AES systems and human markers (Bridgeman, Trapani, & Attali, 2012; Burstein & Chodorow, 1999, 2010; Burstein *et al.*, 2003; Landauer, Laham, & Foltz, 2003; Powers, Burstein, Chodorow, Fowles, & Kukich, 2001)

On the other hand, there has been and still is significant opposition to AES, particularly to the idea, originated by Page, that 'it might replace human scoring' (Ericsson & Haswell, 2006; Herrington & Moran, 2012; Perelman, 2012). Harsh criticism comes particularly from the community of writing researchers. The major organisation *Conference for College Composition* for writing researchers has actively opposed AES during the last decade. Writing professionals claim that such systems prepare their students to write for machines, writers writing to computers (Herrington & Moran, 2001), and therefore they

say: “Because all writing is social, all writing should have human readers, regardless of the purpose of the writing … We oppose the use of machine-scored writing in the assessment of writing” (Deane, 2013, p. 8). They have not revised their statement yet, although there has been a great deal of AES deployment over the last 15 years. Critics argued that the replacement of human markers by a machine would not just threaten the jobs of tutors, but also change students’ sense of what it means to write in school and university (Herrington & Moran, 2001).

Common criticisms of AES (based on (Cheville, 2004; Ericsson & Haswell, 2006)) focus on the capability of such systems to interpret meaning, evaluate factual correctness of the content, and quality of the argumentation. Machines cannot truly read, understand an essay and interpret its meaning (Attali, 2013). Therefore, there is a possibility that such systems can be gamed as AES systems can be insensitive to particular features in student writing that human markers might detect and penalise, such as repetition and lack of coherence (Deane, 2013). There is little research regarding the impact of AES on writers’ behaviour, or on the view of it as a barrier to be gamed and manipulated by tricks rather than as a person to communicate with (Deane, 2013). The biggest opposition to AES focuses on when it is deployed as a replacement for a human scorer, when it becomes the sole scorer. However, such an extreme situation is rare as even the widely-known ETS systems use AES as a complement to the human marker.

It is true that current AES systems do not mimic human markers’ ability to measure conceptual reasoning, thus AES measures a narrower range of skills than human markers (Deane, 2012), though they could measure a lot that human markers do not pay attention to. Such systems therefore are criticised as they fail to measure higher-order writing skills such as high-quality and strong argumentation due to their limited nature (Attali, 2013). For example, the E-rater measures efficiency in ‘knowledge-telling’ writing and cannot score the ‘knowledge-transforming’ writing well enough. In his research, Bennett (2011)

reports on the use of AES in persuasive writing style. He concludes that although the overall correlation between human and machine scores are high, AES systems are better at scoring essays which are marked based on a text-production rubric that values fluency, effective word choice, and accuracy of the text production than they are at scoring essays which are marked based on a critical-thinking rubric that values effective argumentation and attention to the audience. When the focus of assessment is on students who need practice to improve their fluency, and control their text production processes with less cognitive load, the capacity of the AES is relatively strong (Kellogg & Raulerson, 2007); but if the focus is on quality of argumentation, AES is relatively weak (Deane, 2013). Therefore, it is not reasonable to deploy AES as the sole scorer. Instead, it can be deployed in combination with human markers instead.

Attali (2013) pointed out that there is a lack of understanding of what human markers do in their evaluation. He mentioned that the primary goal of AES is to ensure that human markers think similarly about what constitutes high or low quality student writing so that machine scores measure the same elements as human markers. However, there is evidence showing discrepancies between the way human markers interpret the quality of the same essay (Attali, Lewis, & Steier, 2012). For instance, 'rater severity/leniency', the systematic assignment of lower or higher ratings than the average of ratings assigned by other markers, is one of the main discrepancies between markers (Engelhard & Myford, 2003). However, even 'rater calibration methods', extensive training prior to marking and use of marking rubrics to bring consensus, cannot alter the 'rater severity' (Engelhard & Myford, 2003).

If human markers are inconsistent and unreliable, then the machine cannot be trained effectively (Bridgeman, 2013). Therefore, the aim of mimicking human markers is a difficult task to achieve. Bridgeman (2013) discusses how to assess the rater reliability so that machines can be trained better. However, in order to deploy an AES system by considering such limitations, this deployment must be sensitive to AES' own limitations as

well. It does not understand the essay and therefore it is limited to measuring a subset of the written context; therefore, AES should currently be considered as a "complement to human scoring" (Attali, 2013, p. 194). A "division of labour" approach (Attali, 2013, p. 194) between human markers and machines can be used to overcome such issues.

Unlike the initial intention of Page, AES should be used as a "complement to (instead of replacement for) human scoring, limited in its ability to measure a subset of the writing construct" (Attali, 2013, p. 182). "No assessment technology should be applied blindly; but neither should any method be rejected a priori, without considering how it can be used to support effective learning and teaching" (Deane, 2013, p. 18).

Although, in general, AES systems mimic the human markers well enough that various studies show high correlations, just because the approach works well on average does not guarantee that it will work well in all population subgroups (Bridgeman, 2013). There are several studies (Bridgeman *et al.*, 2012) of how well such systems work with student essays written by people of different gender, race, ethnic, and language backgrounds. However, no studies are available on how automated essay evaluation works between different disciplines and student levels, possibly because the systems that are evaluated are mass-market ETS systems that only work in student essays for entrance exams and which do not differ in level or discipline.

3.5 How does it work? - Automated text analysis techniques

The main approaches in the field of computational linguistics relating to automated text analysis are: "comparing text fragments as bags of words in vector space, using lexical resources and using Latent Semantic Analysis (LSA)" (Gabrilovich & Markovitch, 2007). These techniques are introduced next, and critiqued in relation to the objective of this thesis.

3.5.1 The bag of words approach

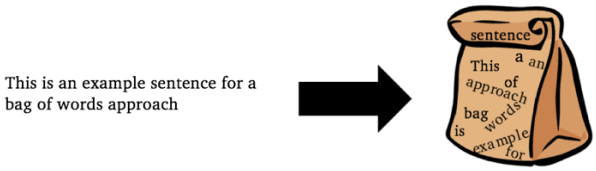
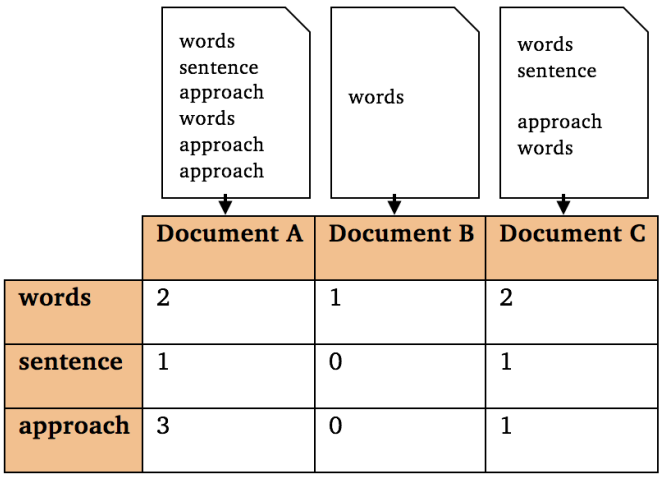


Figure 3.1 Bag of words approach

Text documents can contain thousands of words which provide the starting point for approaches which treat a document as a metaphorical "bag of words" (Figure 3.1).



	Document A	Document B	Document C
words	2	1	2
sentence	1	0	1
approach	3	0	1

Figure 3.2 Bag of words approach (term frequency)

For example, when a user wants to use a query to search for similar documents, the following steps are applied to find out the proximity of a query, how close it is, to a document. As shown in Figure 3.2, the query terms (*words*, *sentence*, *approach*) are searched for in documents (*A*, *B* and *C*). The term frequency is the total number of matches between query terms and occurrences (*words* occurs *twice* in *document A*). A TF analysis would show that documents *A* and *C* are most similar to each other, with document *A* being the best match to the user query.

A strength of this approach is that it scales to millions of documents. However, term frequency does not take the word order into account: the 'bag' contains a jumble of words. Thus a document that claims "John is younger than Mary" is considered identical to a document that claims "Mary is younger than John".

This approach shows the distribution of documents as vectors in a vector space, an algebraic representation of text documents. Similarity or proximity of two documents is calculated as the sum of all the term frequencies. Term frequency (tf) of a given term $_t$ in a document $_d$ is defined ($tf_{t,d}$) as the number of times that $_t$ occurs in $_d$ (see Figure 3.2). Frequency equals the number of counts. So in the example query "*words sentence approach*", similarity is calculated as follows.

- $tf_{words,document\ A}$ equals 2 (as document A contains two '*words'* terms)
- $tf_{sentence,document\ A}$ equals 1 (as document A contains one '*sentence*' term)
- $tf_{approach,document\ A}$ equals 3 (as document A contains three '*approach*' terms)
- So the sum of these three tfs will give the overall proximity of the example query to document A which is 6.
- In contrast, tf would be 1 for document B as it does not contain '*sentence*' or '*approach*'terms and it only contains one '*word*' term (see Figure 3.2).

In terms of its logic, this approach accepts that a document with ten occurrences of the term is more relevant than a document with one occurrence of the term. This might be true; but even if it is, it may not be ten times more relevant. Relevance does not increase proportionally with term frequency. One approach to overcome this problem is by calculating the score as the sum of one plus log of term frequency of each term in a document. This has been shown to return superior results in information retrieval.

An assumption made in this approach is that texts that use synonyms will convey similar messages. Regardless of examining how and why authors used specific words, a bag of words approach takes different texts (sentences or documents), compares them through the

frequency of each word in both texts, and decides whether or not they convey similar messages based on this number. A limitation of this is the presence of homonyms (same term, different meanings): "This is a great book"; "You can book your flights from this website".

Since good academic writing requires the appropriate use of relevant vocabulary from the domain of discourse (Hyland & Tse, 2007; Ivanič, 2004), term frequency is a partial solution. It is hard to imagine a high quality essay which does not cover much of the expected vocabulary, and a bag of words approach will assist in giving feedback on those grounds. However, the absence of any understanding of the role that different words play in language is an important limitation. Approaches that take into account word meanings add greater sophistication, and these are introduced next.

3.5.2 Lexical resource-based approaches

> *"Any system that hopes to process natural languages as people do must have information about words and their meanings*" (Miller, 1995).

It is possible to go beyond exact word-matching by using lexical resources, databases containing machine-readable dictionaries which rapidly search lists of words, making them a very popular natural language processing (NLP) approach (Vossen, 1998). They provide the opportunity to consider various linguistic phenomena such as synonymy, antonymy, and hyponymy. Synonymy is similarity of meaning, in terms of substitutability, it is having the same meaning as another word or phrase in the same language. "The antonym of a word x is sometimes not-x, but not always. For example, 'rich' and 'poor' are antonyms, but to say that someone is not rich does not imply that they must be poor; many people consider themselves neither rich nor poor" (Miller, Beckwith, Fellbaum, Gross, & Miller, 1990, p. 7). "A hyponym inherits all the features of the more generic concept and adds at

least one feature that distinguishes it from its superordinate" (Miller *et al.*, 1990, p. 8). To illustrate hyponym, maple is a hyponym of tree, and tree is a hyponym of plant.

Throughout the years, different ranges of lexical resources have been created for automatic semantic processing of text documents. Each of these lexical databases serves different aims. A widely-known lexical database for the English language is Princeton's WordNet[1] that contains sets of grouped nouns, verbs, adjectives and adverbs (see Table 3.1). These are interlinked by means of conceptual-semantic and lexical relations.

Table 3.1 Semantic Relations in WordNet (Miller, 1995)

Semantic Relation	Syntactic Category	Examples
Synonymy (similar)	N, V, Aj, Av	pipe, tube rise, ascend sad, unhappy rapidly, speedily
Antonymy (opposite)	Aj, Av, (N, V)	wet, dry powerful, powerless friendly, unfriendly rapidly, slowly
Hyponymy (subordinate)	N	sugar maple, maple maple, tree tree, plant
Meronymy (part)	N	brim, hat gin, martini ship, fleet
Troponomy (manner)	V	march, walk whisper, speak
Entailment	V	drive, ride divorce, marry

Note: N = Nouns Aj = Adjectives V = Verbs Av = Adverbs

PropBank (Palmer, Gildea, & Kingsbury, 2005) adds a layer of predicate-argument information, or semantic role labels which are very different to those used by WordNet. PropBank is an annotation of one million words of the Wall Street Journal portion of the Penn Treebank II (Marcus *et al.*, 1994) with predicate-argument structures for verbs, using semantic role labels for each verb argument. It can therefore understand within the same

[1] http://wordnet.princeton.edu/

paragraph or document that the window in the sentence 'John broke the window' is the same window in 'The window broke' (Palmer *et al.*, 2005).

Used for a completely different purpose, SentiWordnet (Esuli & Sebastiani, 2006) is a database used for performing sentiment classification on text documents. This is mainly used for opinion mining which can also be used to extract opinion-bias information from documents.

In short, there are various resources/databases that serve a range of purposes, and each has its own strengths and weaknesses. Although some lexical databases have the same purpose, such as providing synonymy information, their results can vary immensely (see 3.2 for example).

Table 3.2 Subsets of the synonyms provided by different lexical resources for the adjective "bright"

(Sinha & Mihalcea, 2009)

Lexical Resource	**Returned Words**
WordNet	burnished, sunny, shiny, lustrous, undimmed, sunshiny, brilliant
Encarta (Soukhanov, 1999)	clear, optimistic, smart, vivid, dazzling, brainy, lively
Roget (Roget, 1911)	ablaze, aglow, alight, argent, auroral, beaming, blazing, brilliant
TransGraph	nimble, ringing, fine, aglow, keen, glad, light, picturesque
Lia	red, yellow, orange, pink, blue, brilliant, green, white, dark

In order to overcome this limitation, a common strategy is to combine the benefits of multiple lexical resources, which is referred to as lexical substitution, e.g. (Loper, Yi, & Palmer, 2007; Sinha & Mihalcea, 2009). This method was announced in the workshop on semantic evaluations (SemEval[2]) in 2007. The idea is for the selected target (such as the word 'bright') to return synonyms from several lexical resources, as in Table 3.2. The appropriate ones can then be selected (Sinha & Mihalcea, 2009).

[2] http://nlp.cs.swarthmore.edu/semeval/

This approach is most appropriate for individual word comparison although some resources allow predicate-argument structure. Therefore, its adaptation for comparing longer texts, sentences or paragraphs requires an extra level (Gabrilovich & Markovitch, 2007) and consequently, it is not sufficient as an approach for conveying the rhetorical level of comments and arguments required in academic writing.

3.5.3 Latent Semantic Analysis (LSA)

The third approach, Latent Semantic Analysis (LSA), is a fully automatic, mathematical and statistical technique for automatic indexing and retrieval. It was designed to overcome the problems of other retrieval techniques that try to match words within user queries with the words of the document. As discussed in previous sections, a key deficiency of these kinds of information retrieval techniques is that individual words may not be reliable enough to retrieve the conceptual content. This is because there are various ways of expressing any given concept; "the literal terms in a user's query may not match those of a relevant document" (Landauer, Foltz, & Laham, 1998). LSA was designed to overcome this term-matching retrieval problem.

LSA assumes that there is some underlying latent semantic similarity between the user query and the documents. For instance, two documents might be semantically similar even if they do not contain the same words: "the words searchers use often are not the same as those by which the information they seek has been indexed" (Landauer *et al.*, 1998). The fundamental deficiencies of most information retrieval techniques relate to three issues (Landauer *et al.*, 1998):

- Synonymy: There are many ways to refer to the same object. Users in different contexts or with different needs, knowledge, and linguistic habits describe the same information by using different terms. For instance, there is only a 20% possibility that two people choose the same main key word for a single well-known object.

- Polysemy: Most words have more than one distinct meaning. In different contexts or when used by different people, the same term can take on a different meaning.

- Inadequate number of index terms: The index terms identified for comparing two documents, or the user query with documents are incomplete. They only contain a fraction of the terms under which users will try to look them up. The documents themselves do not contain all the terms that users will apply in their queries. For example, a writer might use the words 'access' or 'retrieval'; but a user might use the word "look-up" instead.

As an illustration, a user might want to search a dataset with 100 documents using the query "information look-up" to retrieve any related terms. If the query includes the compound word, it could also have contained words with similar meanings like "access" or "retrieval" as well. Therefore, documents containing these two words can be returned to the user even though these terms were not part of the original query. LSA's job is to find ways to predict what terms are really implied by a query or apply to a document; in other words, it is designed to work out the latent semantics within a query.

LSA does not depend on word frequency, syntax, or pragmatics to measure semantic similarity between text samples. To determine the similarity, LSA employs a mathematical formula known as singular value decomposition (Crossley, McCarthy, Salsbury, & McNamara, 2008). Unlike traditional NLP and artificial intelligence programs, it does not use human-constructed dictionaries such as lexical resources, knowledge bases, semantic networks or grammars; it only takes raw text input parsed into words and separated into meaningful passages such as sentences or paragraphs (Landauer *et al.*, 1998).

The LSA can be used for automating the marking. To do this, LSA needs to be trained in respect of domain-representative text (Foltz, Laham, & Landauer, 1999). It needs a "semantic space" which has been trained with the representative text, so for example if the system will be used to mark biology essays, then it will be trained with a relevant biology

textbook. In the LSA, the essay to be graded is compared to all other essays and text within this semantic space, and the grades of similar essays are then used to predict what grade the expert would have given. One example of a system that uses LSA is the Intelligent Essay Assessor (IEA), a set of software tools for scoring the quality of essay content.

3.6 Automated feedback on student writing

The systems explained in section 3.3, automated essay scoring systems, focus solely on assessment, rather than feedback (Rivers, Whitelock, Richardson, Field, & Pulman, 2014). Automated essay evaluation technologies, however, can be used not just for speedy scoring, but also for providing students with feedback which is specific to their writing in order to help them improve their writing skills. Nowadays, various technologies exist that provide automated feedback on students' writing. For instance:

- *OpenEssayist*, a web application system, has been designed to help students in higher education understand the strengths and weaknesses in their draft essays. There are two components to the system, the learning analytics engine, *EssayAnalyser* which is a summarisation engine, and the web application that provides feedback for students, *OpenEssayist*. *OpenEssayist* processes essays and offers feedback through key phrase extraction, by identifying which phrases are most suggestive of the content, and extractive summarisation which identifies key sentences (Whitelock, Twiner, Richardson, Field, & Pulman, 2015). Each essay is automatically pre-processed using modules from the Natural Language Processing Toolkit (Bird, Klein, & Loper, 2009), that typically uses large bodies of linguistic data, or corpora, (lexical resources as described in section 3.5.2).
  The rationale for developing *OpenEssayist* was based on the knowledge that university students find essay writing to be challenging task. Therefore, a system that provides immediate feedback, or "advice for action"(Whitelock, 2010) on

students' draft essays could be one way of overcoming this challenge. Advice for action enables students to "move forward in their studies by using the information obtained from the analysis" (Whitelock, Field, Pulman, Richardson, & Van Labeke, 2014). *OpenEssayist* invites students to engage with and reflect on their work, in any subject domain (Whitelock *et al.*, 2015). It is for formative assessment, not for giving the students marks, but to improve their work through their understanding of the requirements of academic essay writing (Whitelock *et al.*, 2015).

- *Coh-metrix* is an automated natural language processing tool that looks for 'cohesion' indicators, i.e. how well the written text 'hangs together', including word characteristics, sentence characteristics, and the discourse relationships between ideas in the text (McNamara & Graesser, 2012). "In *CohMetrix*, sentences, paragraphs, and texts are measured as weighted vectors and LSA values" (Crossley *et al.*, 2008). The 'cohesion' of a text refers to the presence or absence of cues in the text that help the reader to understand the relationship between the ideas presented, and 'coherence' is perceived as what the reader takes from it. These cues include words and ideas repeated across sentences and sections, referential overlap, and connective words such as 'because', 'however', 'therefore' (McNamara, Crossley, & Roscoe, 2013).

  McNamara, Crossley, and McCarthy (2010) used *Coh-Metrix* to investigate the role of cohesive devices and linguistic sophistication in explaining human ratings of essay quality. The definition of writing quality has been based on human judgments, expert markers from academic communities, who have been trained to reliability using a standardised marking rubric.

  Researchers explain that when the quality of writing improves, the number of cohesion features does not necessarily increase; there is no indication that higher scored essays were more cohesive. "Higher scored essays were more likely to contain linguistic features associated with text difficulty and sophisticated

language" (McNamara, Crossley, & McCarthy, 2009, p. 73). More advanced readers and writers use cohesion connectives less (McNamara, Graesser, McCarthy, & Cai, 2014), and more skilled writers use more sophisticated language (McNamara *et al.*, 2009).

Coh-Metrix has become the foundation in the Writing Pal intelligent system, which is explained next.

- *Writing Pal* (W-Pal) is an intelligent tutoring system designed to provide writing strategy instruction, game-based strategy practice, and personalised formative feedback for secondary-school and developing first-year undergraduate writers (McNamara et. al, 2013). *W-Pal* provides students with training on the use of strategies to improve their writing quality and, more specifically, on how to write essays. With *W-Pal*, students are provided with lessons on strategies for the various phases of writing, such as generating and organising ideas before writing, drafting an essay with strategies on building the structure of introduction, body, and conclusion, and revising the essay (McNamara *et al.*, 2013). Each of these lessons includes practice in the form of mini-games.
- *LightSide Labs* is an educational technology company, dedicated to improving student writing skills. It has developed a framework which provides automated feedback on student writing in the K-12 classroom (Mayfield & Rosé, 2013). The *LightSide* framework is open source allowing its users to develop new feature extraction and machine-learning technology. *LightSide* is not an LSA application, but instead a machine-learning application that uses samples of graded student writing as input into the algorithms to help the scoring engine learn and train. Although it uses a machine-scoring algorithm, similar to the systems described in section 3.3, it also automatically generates specific and actionable feedback on student writing during the writing process.

The *LightSide*'s *Revision Assistant*, an online program to help school teachers when assessing K-12 students' writing, provides automated writing support to students, on demand, as they draft, with each student's process tracked for teachers to review. The *Revision Assistant* system requires training, by learning how educators grade, evaluate students' work, and provide feedback. This information is then used to provide students with automated feedback throughout the writing process.

- *Turnitin*[3], which is a popular plagiarism checker, used worldwide in various schools and universities. As the *LightSide Revision Assistant* engages students in the writing process by providing them with positive, useful and instantaneous comments about their writing assignments, *Turnitin* has stepped forward by acquiring *LightSide Labs* to build on the algorithms developed by *LightSide Labs*, in order to provide automated feedback as well.
- *Grammarly*[4] is a free web-based application that is a writing-enhancement platform developed for proofreading and plagiarism-detection for any writer.

3.6.1 Summary

The tools explained above analyse student writing using various aspects of written text such as cohesion, grammar, plagiarism, and structure. What has been missing, however, is explicit help and support specifically targeting argumentation within student writing, where this thesis' interest lies. There are argument mapping tools available such as *Cohere* (Buckingham Shum, 2008), and *Compendium* (Okada, Buckingham Shum, & Sherborne, 2008), that help to construct argument in a visual way but these do not assess the quality of writers' argumentation in written text.

[3] http://www.turnitinuk.com/en_gb/login
[4] https://www.grammarly.com/

There are several automated assessment/feedback technologies that are designed to help students: for K12 such as *W-Pal* and *LightSide,* and undergraduates such as *OpenEssayist*. Although there is widespread research on providing feedback to the students themselves about their writing, automated technologies are not only developed for students, but also tutors to provide support in their essay assessment and feedback practices, which is a particular interest of this thesis. To respond to that challenge, *OpenMentor* was developed for tutors to reflect on the quality of their feedback with respect to the mark awarded on electronically-submitted student assignments (Whitelock *et al.*, 2012). The specific focus of *OpenMentor*, however, has been on feedback provided earlier which can then help tutors to understand what constitutes constructive and supportive feedback for future reference. *OpenMentor* goes through the assignments that tutors have already commented on, extracting these comments, and classifying them. It uses pre-determined benchmarks to estimate what 'ideal' distributions of tutor comments should look like, and then displays to the tutors the difference between the actual distribution and what would be considered ideal (Whitelock & Watt, 2007).

The technologies and tools described above are all developed for the analysis of student writing, with the exception of *OpenMentor*; they are more focused on providing feedback to the student him/herself, however, who can reflect on this feedback, revise their drafts and ultimately improve their writing. The focus of this thesis, however, is on ways in which educators and academic tutors can be supported in overcoming any issues related to the process of giving feedback, as discussed earlier. Although *OpenMentor's* focus is on helping tutors with their feedback processes, it does not specifically target its feedback on the automatic analysis of argumentative patterns in students' writing.

As identified in the previous chapter, argumentation is a key requirement for student writing, which is articulated through meta-discourse, linguistic cues which engage the readers, and explicitly convey the authors' intended meaning, expressing their viewpoint,

argument and claim, and signalling their stance (Hyland, 2005). When assessing student writing, academic tutors, as critical readers, look for students' ability to present and pursue well-reasoned and strong arguments through scholarly argumentation. Academic tutors will therefore necessarily be examining meta-discourse in students' writing as signals of the intellectual moves that make their reasoning visible. None of the above technologies has been developed to analyse meta-discourse in student writing automatically, therefore they will not be used for this research. The two approaches explained in the following sections are influential techniques for this thesis, since they are the only two that have been developed specifically for automatic analysis of meta-discourse in writing. Two influential techniques for this thesis, argumentative zoning and the Xerox incremental parser, have both been developed for automatic analysis of meta-discourse, and are explained in the following sections.

3.7 Argumentative Zoning

Argumentative Zoning (AZ) was developed by Teufel (1999) as the first attempt to annotate rhetorical moves in research articles automatically. Rhetorical 'move' refers to "a discoursal or rhetorical unit that performs a coherent communicative function in a written or spoken discourse" (Swales, 2004, p. 228). Accepting Myers (1992) definition of argument: "any proof, demonstration, or reason that is useful for persuading the audience of the validity of a statement", AZ was developed on the premise that "arguing is an important part of presenting an idea" (Teufel, 1999). AZ is an analysis of document structure based on the idea that various rhetorical moves (such as critiquing existing work of others, making a goal statement, etc.) in scientific text documents form a scientific argument.

AZ analysis assumes that rhetorical pieces within the text should be classified based on the ownership of the ideas in the paper (such as new contributions, citable ideas of others, background knowledge that everybody accepts, etc.), and the sentiment towards the cited

work. The ultimate aim of Teufel's work was to provide an intelligent library search tool for researchers that can summarise single or multiple research papers, and display a visual relationship between papers through the use of citation maps.

Argumentative zoning was built on Swales (1990) model of argumentative moves. Swales' model is based on the analysis of journal articles representing a variety of discipline-based writing practices. Swales examined the introductions to 48 articles in the natural and social sciences, and found that most of them contain a sequence of rhetorical 'moves', (*Create a Research Space, CARS*), which have been used to analyse text in a three-move structure.

Move 1: Establishing a territory	Step 1 Claiming centrality and/or Step 2 Making topic generalization(s) and/or Step 3 Reviewing items of previous research
Move 2: Establishing a niche	Step 1A Counter-claiming or Step 1B Indicating a gap or Step 1C Question-raising or Step 1D Continuing a tradition
Move 3: Occupying the niche	Step 1A Outlining purposes or Step 1B Announcing present research Step 2 Announcing principal findings Step 3 Indicating RA structure

Figure 3.3 Swale's CARS Model (RA = Research Article)

Swales (1990) articulated the move analysis, as a representation of academic research articles in terms of hierarchically organised text made up of distinct sections; each section can be subdivided into moves, and each move can be broken down into steps. Based on the figure 3.3 above, the 'introduction' includes three basic moves: move 1 in the beginning, followed by move 2 and concluded by move 3 (Berkenkotter, 1989).

- Move 1: Establishing a territory (establish the field in which the author works)
- Move 2: Establishing a niche (justify the present study by indicating a gap in current knowledge or by raising questions)

- Move 3: Occupying the niche (introduce and describe the present study, own study, by indicating what the investigation that author is reporting will accomplish for the field).

Swales (1990) argued that each of these moves can be made through one or a series of 'steps'. Teufel took Swales' idea as a basis: "[the] argumentative status of a certain move is visible on the surface by linguistic cues", which means authors introduce linguistic cues (meta-discourse signals) while writing (Teufel, 1999, p. 84). These can be identified to understand and interpret the argumentative and rhetorical status of authors' writing and their stance.

In addition to Swales' model, argumentative zoning was built on Hyland's system of the description of meta-discourse. Meta-discourse refers to the features of text that provide linguistic cues which engage the readers, and explicitly convey the authors' intended meaning, expressing their viewpoint, argument and claim, and signalling their stance (Hyland, 2005). Rather than simply defining meta-discourse as 'discourse about discourse', Hyland (2005) defined the concept of 'meta-discourse' as an important element of the document, that is not only used to organise ideas but also to relate to readers. It is an umbrella term that helps to relate the text to its context, which glues the important parts of a text together but, more significantly, it helps readers to understand existing knowledge and strategies used by other members (authors/researchers) of the subject area, as well as the writer's stance towards these.

According to Teufel, the definition of the argumentative zones is given by the single rhetorical act, which are salient sentences. These sentences are landmark sentences that include meta-discourse cues like 'in this paper we develop a method for' or 'in contrast to REFERENCE, our approach uses...'. Teufel's particular interest is in the rhetorical status of these landmark sentences with respect to the communicative function of the whole paper.

"AZ is independent of writing style, subject matter, and, to a certain degree, subdomain, but relies on text type specific expectations (communicative acts)" (Teufel, 1999, p. 22). Teufel's approach takes each research paper to be one rhetorical act. She defined seven categories, argumentative zones (as given in the figure below), which cover an entire article. This model of scientific argumentation is based on the idea that scientific articles have typical argument structures regardless of their discipline, such as expressions of the author's stance towards other work (Teufel, 1999). Therefore, the claim is that they are not specific to a domain, but are discipline-independent, since the theory and technique of AZ has been shown to be robust and operational (Teufel, 1999).

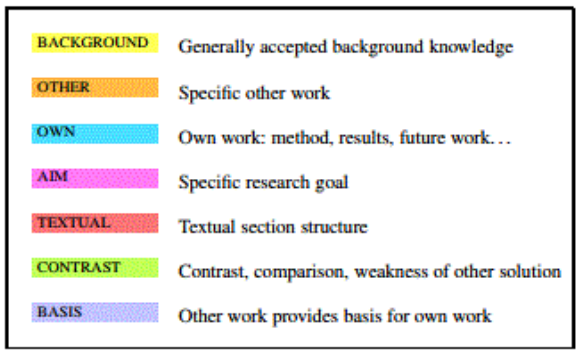


Figure 3.4 Argumentative zones

In her work, Teufel investigated, with a corpus of 200 papers, how humans perform the analysis, and how much they agree or disagree. She found that they agree to a great extent, and how an automatic, rather shallow process can apply the analysis, based on machine learning and features of sentences. The ultimate aim of her work was to provide an intelligent library search tool for researchers that could include the summarisation of single or multiple articles and also improved citation indexes, by means of citation maps which could help people grasp relationships between papers.

Originally, argumentative zoning was proposed for automatic summarisation and information retrieval tasks. Later, it was also used for educational purposes (Feltrim, Teufel, das Nunes, & Aluísio, 2006) and citation indexing (Teufel, 2006). Since the theory

and technique of argumentative zoning are shown to be robust and operational, subsequent work consisted of annotation experiments in different disciplines, including chemistry (Teufel, Siddharthan, & Batchelor, 2009) and biology (Mizuta, Korhonen, Mullen, & Collier, 2006).

AZ has become an influential approach to the automated summarisation of scientific articles that has been built upon by the Xerox Incremental Parser, as explained in the following section.

3.8 Xerox Incremental Parser (XIP)

3.8.1 Background – the XIP syntactic parser

Parsing is the act of splitting up information into components. In computing, particularly in linguistics, parsing refers to the process of analysing text into its constituents, logical syntactic components. Robustness, the ability for a language analyser to provide useful analyses of real-world input text such as news, e-mails, and articles, is a key issue for natural language processing especially for parsing (Aït-Mokhtar *et al.*, 2002).
Since widespread, shallow and partial parsers do not allow deeper language analysis, the Xerox Incremental Parser (XIP syntactic parser) has been developed by Xerox Research Centre Europe in order to build a robust analyser that tackles deeper linguistic phenomena (Aït-Mokhtar *et al.*, 2002). It is called 'incremental' due to its systematic incremental methodology that allows the analyser to "go beyond shallow parsing to deeper language analysis, while preserving robustness" (Aït-Mokhtar *et al.*, 2002, p. 1). 'Incrementality' is a methodological principle used to build robust parsers that rely on computationally tractable syntactic descriptions for parsing to a deeper level of language analysis (Aït-Mokhtar *et al.*, 2002). The XIP syntactic parser refers to a generic, rule-based, robust deep system based on such methodology.

3.8.2 The XIP rhetorical parser

The generic XIP syntactic parser system enables robust deep parsers to be built. The component of particular interest to this thesis is the XIP's rhetorical parser, which has been developed and implemented on top of the generic syntactic parser. At first, the XIP rhetorical parser, referred to as the XIP, XIP or the XIP tool interchangeably for the remainder of this thesis, has been implemented to help researchers to keep up to date in their research domain within the growing number of electronic research publications and scientific articles (Sándor, Kaplan, & Rondeau, 2006).

Scientific articles are highly structured and follow argumentative patterns that guide the reader about authors' intention, approach and their thoughts (Hyland, 2005; Ravelli & Ellis, 2005; Sándor *et al.*, 2006; Teufel & Moens, 2002). Readers grasp and comprehend authors' thoughts and stance through these argumentative patterns, which are articulated by *meta-discourse*.

The XIP shares the basic assumption of Teufel's argumentative zoning (AZ), namely, that rhetorical moves can be detected from the author's language use. Teufel's work introduced a new rhetorical-level analysis of scientific research articles. However, XIP takes a different approach to the rhetorical analysis of scientific articles. Instead of covering the whole article, this approach aims to highlight the main research issues that the article handles. Rather than seeing the whole paper as one rhetorical act, XIP assumes there is more than one rhetorical act within a paper and that these are sentence-specific. Unlike AZ, whose methods work at document level, XIP uses methods that work at sentence level and this is one prominent difference between AZ and XIP analysis.

The XIP detects and labels rhetorically salient, key, sentences in scholarly writing based on the automatic identification of *meta-discourse* conveying the author's rhetorical strategy. It

marks particular discourse function types of sentences in order to provide the reader with additional support for representing scientific work in a structured way.

The first build of the XIP has been implemented for processing biomedical literature in the *PubMed*[5] repository (Lisacek, Chichester, Kaplan, & Sandor, 2005). Ultimately, the XIP rhetorical parser was inspired by Kuhn's (2012) concept of 'paradigm shift' as a basic element of scientific revolution. Lisacek *et al.* (2005) suggested detecting linguistic indicators of 'paradigm shift' in biomedicine research paper abstracts on Alzheimer research as a new text-mining strategy (Lisacek *et al.*, 2005). Lisacek *et al*. (2005) wanted to obtain access to the few papers, among tens of thousands, which propose a new model, point out important contradictions in existing models, pose intriguing questions, and potentially lead to a 'paradigm shift' (Lisacek *et al*., 2005).

To begin with, the XIP developers started with six scientific research abstracts, and built a preliminary model whose performance has since been gradually improved. Based on the analysis of these abstracts, the XIP developers identified the discourse function types that would represent scientific argumentation in biomedical research articles. These were listed as: "background knowledge, logical contradiction, an element insufficiently or not known, research trend, summary sentence, contrast with past findings and substantially new finding" (Sándor *et al.*, 2006, p. 2).

3.8.3 The XIP categories

In the first build, the user (a researcher who searches for scientific articles in *PubMed*) enters a *PubMed* query and an additional list of important keywords that is used for relevance ranking. The output is the list of retrieved abstracts ranked according to the frequency of the desired keywords, and sentences containing the above-mentioned discourse function types highlighted by the XIP (Sándor *et al.*, 2006).

---

[5] PubMed comprises more than 26 million citations for biomedical literature from MEDLINE, life science journals, and online books. https://www.ncbi.nlm.nih.gov/pubmed

XIP's choice of rhetorical moves, referred to as the XIP categories in this thesis, is motivated by various considerations. SUMARY and BACKGROUND KNOWLEDGE relate to conveying main ideas in a straightforward way in the rhetorical construction of research articles. The other categories (CONTRAST, NOVELTY, EMPHASIS, SURPRISE, OPEN QUESTION and TENDENCY) have their roots in Kuhn's (2012) view of science as primarily a problem-solving activity. According to this view, the main ideas found in sentences where the research issues are described fulfil rhetorical functions of contesting, questioning or emphasizing research-related ideas, facts or theories as being significant or new research-related ideas, facts, or theories, of indicating a gap in knowledge, or of pointing out any flaw or contrast related to the research topic.

XIP highlights meta-discourse which conveys the author's rhetorical strategy and annotates the rhetorical functions. The rhetorical functions detected by XIP partly overlap with Teufel's AZ, and are also different from them. The main difference is that the contrasts among ideas are not approached from the viewpoint of intellectual ownership, but rather from the viewpoint of the various ways in which contrasting ideas are introduced. All of the XIP categories have been created based on the linguistic analysis of the experienced researchers' writing in the biomedical literature.The XIP rhetorical parser is rule-based, however, so the concept patterns are identified through rules, which can be modified and revised for different literature. For instance, recently the XIP rhetorical parser has been tested on law literature (Knight, Buckingham Shum, Ryan, Sándor, & Wang, Forthcoming). It should be noted here that the XIP tool is not in active development, but it has been versioned based on the needs of different research domains, discourses and types of writing, due to its rule-based algorithm, as explained in the next section.

3.8.4 The XIP algorithm and labelling sentences

The XIP is not underpinned by a specific linguistic theory. Even though there is no obvious linguistic theory linked to XIP, it is developed based on empirical evidence and linguistic concepts. The development of empirical evidence was incremental, meaning it was not built on a fixed development corpus.

Although the underpinning theory is not linked to an existing linguistic theory, some theoretical works such as scholarly rhetorics analysed in genre studies (Swales, 1990), the study of *meta-discourse* (Hyland, 2005), and discourse analysis (Lewin, Fine, & Young, 2005) have links to the XIP's *concept-matching* algorithm. Ait Saidi (2015) recently linked concept-matching to these linguistically motivated analysis frameworks, like meta-discourse and phraseology studies; however, this research has not proposed linking it to an existing linguistic theory (Ait Saidi, 2015).

Developers of the XIP tool proposed the detection of 'paradigm shift' through a conceptual analysis, which is called 'concept matching'. The concept-matching grammars are built on top of a general, rule-based, robust dependency grammar, the XIP syntactic parser (Sándor *et al.*, 2006). The concept matching framework has subsequently been used to model the labelling of salient sentence types as SUMMARY, BACKGROUND, CONTRAST, NOVELTY, EMPHASIS, SURPRISE, OPEN QUESTION and TENDENCY, as described in the following table.

Table 3.3 Rhetorical Functions identified by XIP

SUMMARY	summarising the goals or results of the article
BACKGROUND	describing background knowledge necessary for understanding the article's contribution
CONTRAST	describing tensions, contrasts between ideas, models or research directions
NOVELTY	description of new ideas, conveying that an idea is new
EMPHASIS	emphasising the importance of ideas
SURPRISE	descriptions of ideas as being surprising
OPEN QUESTION	describing problems that have not been solved
TENDENCY	description of research trends, emerging research directions

Concept matching is a sort of pattern matching, which is often used in automated linguistic analysis. XIP uses a concept-matching method which aims to detect particular rhetorical formulations in text documents. These are neither propositional (like the predicate-argument structure discussed in the *PropBank* lexical dictionary example) nor thematically strongly related. Rather, these formulas add categorisations to propositions which indicate the rhetorical function of the information conveyed by the propositions (Sándor, 2006). For instance, SUMMARY is a function by which the author can refer to the issues dealt with in (parts of) the text. In the following examples, the parts of the sentence carrying out the rhetorical function of SUMMARY are in italics; these are the meta-discourse cues that convey SUMMARY statements:

The purpose of this article is to develop the idea that the procedures in any given classroom or laboratory exercise should be definitely determined by the specific aim, which the instructor has in mind to accomplish.

The perspective I shall use in this essay relies heavily on the view of professionalisation presented in Andrew Abbott's brilliant study, The System of Professions (Abbott, 1988).

This paper explores social practices of propagating 'memes' (pronounced, 'meems') as a dimension of cultural production and transmission within Internet environments.

The following figure illustrates rhetorical functions classified by the XIP based on some example meta-discourse cues that authors might use.

BACKGROUND
Recent studies indicate…
The previously proposed…
… is universally accepted.

NOVELTY
New insights provide direct evidence…
… suggest a new approach …
Results define a novel role…

SUMMARY
The goal of this study…
Here, we show…
Our results … indicate

EMPHASIS
… is crucial for understanding
valuable information … from
… have provided important advances

TENDENCY
… emerging as a promising approach
Our understanding … has grown exponentially …
Growing recognition of the importance…

SURPRISE
We have identified … unusual
The recent discovery … suggests intriguing roles
We have recently observed … surprisingly

OPEN QUESTION
Little is known
… role … has been elusive
Current data is insufficient

CONTRAST
in contrast with previous hypothesis …
… inconsistent with past findings …

Figure 3.5 Example meta-discourse cues within sentences for each XIP category

The concept-matching methodology that XIP uses assigns a common representation to diverse expressions that convey rhetorical comments. The concept-matching method that XIP uses consists of the use of lexical resources for each constituent concept and co-occurrence rules. To identify the meta-discourse cues that signal a given rhetorical move

by the author, XIP specifies a set of dependencies, co-occurrence rules, in which a specific set of concepts must co-occur, in any sequence.

The XIP rhetorical parser has been validated in the life sciences. However, the lexicons can be optimised for specific domains since discourse communities have their own particular ways of signalling rhetorical moves. As mentioned earlier, the XIP itself has never been updated, but the rhetorical parser's rules can be updated based on the needs of different literature. For instance, developers can add new ways of representing CONTRAST with different sets of meta-discourse cues and word pairings. The next section exemplifies how the XIP rhetorical parser evolved, and why its categories have remained the same for different projects other than biomedical research articles.

3.8.5 The XIP versioning

The original XIP rhetorical parser was developed for the life sciences. The XIP has been evolved and re-versioned for educational science literature, during a collaboration between the XIP developers and The Open University's Knowledge Media Institute for investigating the overlaps and complementarities between the outputs from human analysts and the XIP (De Liddo, Sándor, & Buckingham Shum, 2012).

These overlaps and complementarities were investigated through a human and machine annotation study. For this, by keeping the same rhetorical labelling, the categories as described above, the XIP developers added looser rules to the XIP's algorithm for catching meta-discourse cues specific to educational science literature. The XIP was then used for the automatic annotation of 125 documents, and five human analysts also independently annotated the same documents.

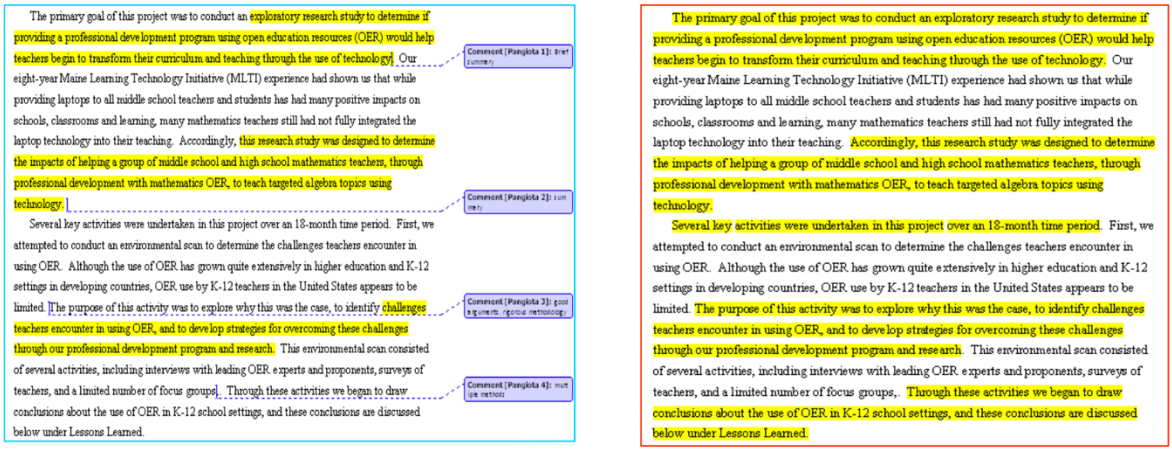


Figure 3.6 Human vs. machine annotation

The figure above demonstrates an example comparison between the human annotation, document on the left and the XIP's output, on the right. De Liddo et al. (2012) then imported these outputs into *Cohere* (Buckingham Shum, 2008) to explore ways of visualising the combined human and machine output. This study exemplifies how the XIP rhetorical parser can be versioned and used for different writing types and needs. However, it should be noted here that although the XIP can be versioned, the lexicons of the XIP tool can only be changed by its developers due to copyright issues.

3.9 Gaps in the literature

Chapters two and three located this thesis firmly within the existing literature, by reviewing the previous theoretical and empirical work on which it is based. The literature review has been divided into two aspects: academic writing and automated assessment of writing. The former investigated what academic writing is, what really matters in student academic writing, what educators think good undergraduate writing is, and how this differs within specific disciplines. The latter provided detailed accounts of key approaches to automated text analysis.

The first chapter of the literature review on academic writing established that academic writing is an important skill that undergraduates should acquire. Since undergraduates move from a school culture to an academic culture, they can find undergraduate writing

difficult and sometimes struggle to understand the norms of university writing; it is difficult to grasp what is expected from them. Thus, there is a mismatch between the expectations of tutors and students with regard to the written assignment. Like students, academic tutors can experience problems in providing timely, effective examples and feedback to their students; therefore, they also require support which could be provided through automated means.

Academic staff evaluate student writing based on their prior conceptions and disciplinary backgrounds. Undergraduate students, however, study and often practise academic writing in modularised degree courses where interdisciplinary groups come together. Therefore, this requires further examination to establish a baseline on academic tutors' views on what makes good student academic writing, and how tutors assess student writing, before proceeding to the use of automated essay assessment. The primary goal of automated systems is to ensure that human markers think similarly about what constitutes high or low quality student writing so that the systems measure the same elements as human markers. Therefore, it is essential to understand what academic tutors value and specifically look for in essay marking.

Academic writing includes similar structural elements in all disciplines, but these expectations are also influenced by subject area. Academics from different communities or cultures may have different perspectives on what constitutes good writing (McNamara, Crossley, & McCarthy, 2010). Thus, although generalisations are made regarding the quality of writing and conventions of academic writing, these judgments may differ across writing tasks, communities, and cultures. The discipline-specificity aspect of academic writing, discussed in section 2.6, is also essential for the meta-discourse cues. As confirmed by (Hyland, 1998) there are variations in meta-discourse between domains. Therefore, this issue needs to be addressed and further explored, how automated support

through automatic identification of meta-discourse can be used for different disciplines and levels.

Even though studies showed how automated essay assessment systems can be affected by differences in gender, language, and country, there are no studies available on how automated essay evaluation works between different disciplines, both in hard and soft disciplines and at different student levels. This is possibly because the systems that are evaluated are the mass-market ETS systems that only work in student essays for entrance exams which do not differ in level or discipline.

There is an unresolved tension felt by human markers towards the idea of automated text analysis, as given in section 3.4. As pointed out during the literature review, "no assessment technology should be applied blindly; but neither should any method be rejected a priori, without considering how it can be used to support effective learning and teaching" (Deane, 2012). Considering the tensions of human markers towards automated analysis, this thesis focusses on this unresolved aspect to evaluate and value both parties' arguments, and to suggest how automated assessment should be effectively deployed.

This thesis will investigate ways in which a specific computational language technology can be used to support educators' essay assessment practices in different disciplines and levels, how educators define the quality, and to what extent this can be captured by this specific tool, so that its output can be presented to tutors to support their assessment practices. Therefore, the generic research questions could be formulated as below.

To what extent can the automated rhetorical parser technology be used to identify indicators of good academic writing in undergraduate student essays and to support educators' feedback processes? Developing from this main research question, the subsidiary research questions are:

RQ 1: To what extent can the [*specific computational language tool*] be used to identify indicators of good academic writing in undergraduate student essays from different disciplines, as judged by the essay grade?

RQ 2: How do educators define the attributes of good student writing and to what degree can the [*specific computational language tool*] identify the presence of these attributes?

RQ 3: How congruent is the [*specific computational language tool*] analysis of student essays with educators' judgement of quality?

RQ 4: In what ways should the output of the [*specific computational language tool*] analysis of student essays be delivered to educators so that they make use of this output to give feedback on student essays?

The next section discusses which specific computational language tool will be used, in order to answer these research questions, and why it is selected as an exemplar tool.

3.9.1 Why XIP?

In section 3.6 some example automated text analysis technologies and tools were introduced. These have been developed for the analysis of student writing and to provide feedback and eventually improve the writing of K-12 and/or undergraduate students in various aspects of their writing such as cohesion, grammar, plagiarism, and structure. These tools will not be used in this research to answer the questions emerging from the literature. As noted in section 3.6.1, the reasons for not using these tools are:

1. The analysis of these tools is not specifically around argumentation, but rather on aspects such as cohesion, grammar, plagiarism, and structure.
2. They all have been developed for the analysis of student writing with an intention of providing feedback to the students themselves. However, the focus of this thesis is to investigate how computational language technologies can be used to support educators in their essay assessment processes.

3. Although *OpenMentor's* focus is on helping tutors with their feedback and assessment practices, it does not specifically target feedback around the argumentative patterns in students' writing.
4. Argument is a key requirement in undergraduate student writing which is articulated through meta-discourse markers. Since these tools have not been designed and therefore targeted to the analysis of such linguistic features, none of these are going to be used as an exemplar computational language technology to answer the research questions of this thesis.

Two influential techniques for this thesis, argumentative zoning and the XIP, are discussed in sections 3.7 and 3.8 respectively. These are the only two that have been developed specifically for the automatic analysis of meta-discourse in writing. Although these two have not targeted student writing specifically, they were both originally developed through the analysis of experienced researchers' articles and journal papers. Although journal writing is not the same genre as student writing, experienced researchers review the literature, articulate arguments based on supported evidence, and present new research by critiquing existing research. This thesis therefore begins to work on the assumption that if salience in journal papers can be extracted through these technologies, the same should be the case with student writing.

The XIP and argumentative zoning (AZ) share the basic assumption, namely, that rhetorical moves can be detected from the author's language use. The work with AZ introduced a new rhetorical-level analysis of scientific research articles. The XIP has been built based on the idea of AZ and moved AZ's assumption forward. XIP takes a different approach to the rhetorical analysis and instead of covering the whole article, it aims highlight the main research issues that the article handles. Rather than seeing the whole paper as one rhetorical act, XIP assumes there is more than one rhetorical act within a paper and that these are sentence-specific. Unlike AZ, in which methods work at document

level, XIP uses methods that work at sentence level; this is one prominent difference between AZ and XIP analysis.

Since academic tutors look at several aspects when assessing student writing, unlike the AZ's document-level method that assigns rhetorical move to the whole text, the XIP's sentence-level analysis method seems more suitable. Educators assess student writing on several aspects. Therefore, XIP's sentence-specific methods that analyse meta-discourse cues for each sentence and label each sentence individually based on the rhetorical meaning conveyed with these cues is preferred. When each category was examined for both AZ and the XIP, AZ's 'OWN' rhetorical label did not seem suitable for student-level rhetorical moves. AZ's OWN move is used for sentences that demonstrate the author's own specific work, a description of its methods, and future work. When considering undergraduate writing, students are not expected to produce a brand new research work, especially in their early years. Therefore, AZ's rhetorical moves cannot be easily mapped against undergraduate writing moves.

On the other hand, with a key focus being on higher-order writing skills such as argumentation, the following points show how XIP categories can map against the ways in which students can argue.

- One can refute an argument and can be against a particular position (CONTRAST).
- One can present important and strong arguments by describing ideas as significant and unexpected (EMPHASIS, SURPRISE).
- One can present an argument through the description of new ideas (NOVELTY) and pursue it through summarising the claims (SUMMARY).
- One can use examples and evidence to support or prove an argument through the description of research trends and the description of previous ideas (BACKGROUND).

- One can indicate a gap with the description of unresolved ideas (OPEN QUESTION).

XIP has been designed to work on peer-reviewed academic research writing, but by a team with no training in education, or intent for it to be used in education. As shown above, however, it connects with education to the degree that there seems to be an overlap in the hallmarks of research articles, the rhetorical moves, and the kinds of argumentative writing that academics seek to nurture in undergraduate students.

Finally, the overall intention is to use an automated analysis of student writing to support academic tutors' essay assessment processes. The human and machine annotation study, explained in section 3.8.5, exemplifies how the XIP rhetorical parser can be versioned and used for different writing types and needs. Therefore, it creates the potential to take this study further with student writing, and to adapt the XIP for the machine analysis of student writing and to perform similar activity with human markers, academic tutors, and annotations.

Answering the generic questions identified in the earlier section requires a specific tool, and the XIP is the best choice for the reasons discussed above. Therefore, this thesis aims to test XIP's automated rhetorical analysis technology within a higher educational context, with an aim of identifying good attributes of scholarly writing across disciplines. This PhD thesis builds the argument that XIP can identify and extract good attributes of academic writing. This claim requires validation which may be conducted through the following research questions. Therefore, the research questions are revised as follows:

RQ 1: To what extent can the automated rhetorical parser XIP be used to identify indicators of good academic writing in undergraduate student essays from different disciplines, as judged by the essay grade?

RQ 2: How do educators define the attributes of good student writing and to what degree can the automated rhetorical parser XIP identify the presence of these attributes?

RQ 3: How congruent is the XIP analysis of student essays with educators' judgement of quality?

RQ 4: In what ways should the output of the XIP analysis of student essays be delivered to educators so that they make use of this output to give feedback on student essays?

The next chapter provides the research methodology undertaken in order to answer these questions.

[This page intentionally left blank.]

4

METHODOLOGY

4.1 Introduction

This chapter describes the rationale for the overall research design, together with the details of the methods employed in this research. Firstly, the primary and subsidiary research questions are presented in section 4.2. Secondly, the research philosophy is discussed in Section 4.3. Following this, the mixed-methods approach that is adopted in this PhD research is discussed along with a justification for the choice of the methodologies. Section 4.4 presents the overall research design followed by a brief explanation of the data collection methods and the data analysis; these are explained in detail in the subsequent chapters. The research design is presented in Section 4.5 and the chapter ends with section 4.6, which includes a discussion of reliability, validity and generalisability issues.

4.2 Research aim and questions

The aim of this research is to investigate whether computational techniques such as Natural Language Processing (NLP), can identify the attributes of good academic writing in undergraduate student essays automatically; and if this proves possible, how best to feed back actionable analytics to support educators in their essay assessment processes. It also sets out to explore the applicability of such automated support within different disciplines. The investigation will focus on a specific computational language technology, XIP, described in chapter 3.8. In order to gain an understanding of how XIP can be adapted to a

higher education framework with the primary intention of analysing student essays, a better understanding of the advantages, drawbacks and challenges of analysing student essays using the current form of XIP is necessary. This would facilitate a specification of improvements that would open up the possibility of applying it in higher education contexts.

Thus the main research question (MRQ) is: **To what extent can the automated rhetorical parser technology be used to identify indicators of good academic writing in undergraduate student essays and to support educators' feedback processes?**

Developing from this main research question, the subsidiary research questions are:

RQ 1: To what extent can the automated rhetorical parser XIP be used to identify indicators of good academic writing in undergraduate student essays from different disciplines, as judged by the essay grade?

RQ 2: How do educators define the attributes of good student writing and to what degree can the automated rhetorical parser XIP identify the presence of these attributes?

RQ 3: How congruent is the XIP analysis of student essays with educators' judgement of quality?

RQ 4: In what ways should the output of the XIP analysis of student essays be delivered to educators so that they make use of this output to give feedback on student essays?

4.3 Research philosophy

Three major frameworks exist for designing a research study in the discipline of social sciences: quantitative, qualitative and mixed methods approaches. Of these, the quantitative methods were developed first, the qualitative methods were developed following this and have been available for the last four or five decades; and the mixed

methods approaches are relatively new and were developed in the last few decade(s) (Denscombe, 2008).

Quantitative research employs experimental methods and quantitative measures to test hypotheses (Hoepfl, 1997), and emphasises the measurement and analysis of causal relationships between variables (Denzin & Lincoln, 1998). Qualitative research, on the other hand, is "any kind of research that produces findings not arrived at by means of statistical procedures or other means of quantification" (Strauss & Corbin, 1990, p. 17). Although there are qualitative researchers whose studies contain big sample sizes (Symonds & Gorard, 2008), qualitative research typically focuses in depth on relatively small samples, even single cases (n = 1), selected purposefully; on the other hand, quantitative methods typically depend on larger samples selected randomly (Patton, 1990). Some notable strengths of quantitative research include the capacity to examine relationships between variables of interest, to test the research hypothesis and to conduct group comparisons (Castro, Kellison, Boyd, & Kopak, 2010). Qualitative research has the strength of capturing findings in original real-world contexts such as human experiences of emotions, beliefs and behaviours. Some limitations of qualitative research include difficulties in drawing definitive generalisable results and assessing associations across observations (Castro *et al.*, 2010). Some qualitative analysts such as Denzin and Lincoln (1994) argue generalisability, reliability, and validity are not the most appropriate terms in this context with qualitative research. Such analysts argue that qualitative research should move away from the strict scientific definitions of reliability and validity, as restricting qualitative research to these confining definitions means that researchers are unable to report all the data they collect with accuracy (Denzin & Lincoln, 1998; John Lewis, 2009; Wolcott, 2005). On the contrary, pure qualitative research studies are often criticised for their small and unrepresentative samples and their limited opportunities for generalising results (Castro *et al.*, 2010).

As both qualitative and quantitative research methods have strengths and weaknesses, mixed-methods research offers the promise of bridging across both traditions (Haverkamp, Morrow, & Ponterotto, 2005) and integrating the strengths of both methods to produce scientifically reliable and valid analysis.

Considering the strengths and weaknesses of each method, this study adopts the ideology of choosing the most suitable method for answering each research question. However, it should be noted here that single method theorists do not believe that "different data types could benefit each other or that methods other than their own could provide a coherent version of reality" (Symonds & Gorard, 2008, p. 2). Symonds and Gorard (2008, p. 16) who called for the death of mixed methods, noted the following:

> A review of the history of mixed methods reveals its beginnings as a bridge between the qualitative and quantitative research paradigms in the era of the paradigm wars. However, its bid for methodological freedom rests on there being two very separate paradigms to begin with. (p.16)

Boyatzis (1998) introduced the terms *quantiphobe* and *qualiphobe* to define researchers who have a fear or dislike of either quantitative or qualitative methods. Teddlie and Tashakkori (2009) added the term *mixiphobes* as another type of researcher, one who chooses a pure orientation and has a fear of mixed-method approaches; therefore, some researchers who have a purely methodological orientation would not adopt a specific approach in their studies even if it seems to be the most suitable method. If a researcher is either a *qualiphobe* or a *quantiphobe*, it is likely that s/he is also a *mixiphobe*; therefore, s/he would not make such decisions. As Symonds and Gorard (2008, p. 10), promotors of the death of mixed methods, also emphasised, mixed methods would be easily adapted "for those researchers who are philosophically committed to bipolar paradigms anyway, and has very little bearing on how research is and can be conducted".

Each of the approaches adopts different philosophical assumptions about what constitutes a 'knowledge claim'. A "knowledge claim means that researchers start a project with certain assumptions about how they will learn and what they will learn during their inquiry"

(Creswell, Plano Clark, Gutmann, & Hanson, 2003, p. 7). Knowledge claims can be related to four different philosophical ideas:

- **post-positivist**: knowledge claims accept that absolute truth can never be found and therefore the evidence established in research is always imperfect and fallible
- **socially constructed**: knowledge claims are made by researchers who seek to understand the context or setting of the participants by engaging with the world they are interpreting through visiting the context or gathering information personally
- **advocacy/participatory**: knowledge claims are made by researchers who believe post-positivist assumptions impose structural theories that do not fit marginalised individuals or groups
- **pragmatic:** knowledge claims are derived from pluralistic approaches used for the problem (Creswell, 2013).

Usually, quantitative researchers accept the post-positivist knowledge claims, qualitative researchers accept the socially constructed knowledge claims and mixed methods researchers accept the pragmatic knowledge claims.

Although experienced researchers often tend to follow one research direction and describe themselves as either qualitative or quantitative researchers, this was not the case for me. As a graduate novice research student without prior pure philosophical orientation, I have not been "imprisoned within one of these purported paradigms" (Gorard & Taylor, 2004, p. 149), and I have not rejected any one philosophical direction completely, which is why I am taking a pragmatic approach.

In this thesis, a mixed-method approach is used to guide the research design in order to benefit from the strengths of both methods and to overcome their limitations. Studies are

considered to be ‘mixed’ if they adopt quantitative or qualitative approaches in one or more of the following ways (Creswell, 2013):

- two types of research question necessitating the use of both quantitative and qualitative approaches
- two types of data collection procedure, for example surveys and focus groups
- two types of data, for example numerical and textual
- two types of data analysis, for example statistical and thematic.

In the first issue of the Journal of Mixed Methods Research, Tashakkori and Creswell (2007) emphasised that it is necessary to distinguish between mixed methods as the collection and analysis of two types of data, and as the integration of two approaches to research. Although the two seem interchangeable, the former is more focused on methods, and the latter on methodology. “Specific research methods are determined by the overall methodological orientation of the researchers” (Teddlie & Tashakkori, 2009, p. 21). Since there is not a specific methodological orientation as a researcher, this thesis adopts the latter interpretation of mixed methods: integration of quantitative and qualitative approaches to research. Therefore, the pragmatic knowledge claims are accepted and a variety of approaches to the collection and analysis of data were considered.

Since the mixed-methods approach is used in this research, pragmatic knowledge claims are embraced. “Pragmatism is not committed to any one system of philosophy and reality” (Creswell, 2013, p. 12). Instead, pragmatists have a freedom of choice – to choose the methods, techniques, and procedures that best meet their purposes (Creswell, 2012). Pragmatists look at ‘what works’ and aim to use the best research method(s) to answer the research question as fully as possible, and researchers are free to choose the methods that best meet their needs. Mixed-methods researchers therefore employ many alternative approaches to the collection and analysis of data rather than adopting one approach.

Researchers like Symonds and Gorard (2008, pp. 10, 16) argue that some of the mixed-methods research studies do not actually meet the standards of mixed-methods research criteria, or fit the mixed-methods research descriptions. Adopting a pragmatic approach in this research allowed for careful consideration of which approach was most appropriate for each research question, rather than accepting one approach straight away. This is in line with mixed-methods research. Therefore, it was necessary to select appropriate datasets for answering research questions and to identify most suitable methods. The next section briefly describes the selection of datasets and considers the strengths and weaknesses of each method.

4.4 Research methods

Detailed explanation of the methodologies considered for each study and the reason for why specific methods were selected rather than another is described in the following chapters. The reasons for choosing both qualitative and quantitative methods can be briefly explained in the context of each research question as follows.

RQ1: *To what extent can the automated rhetorical parser XIP be used to identify indicators of good academic writing in undergraduate student essays from different disciplines, as judged by the essay grade?*

As XIP was originally developed for the analysis of experienced researchers' journal articles in the discipline of science, the first sub-research question explores how XIP performs in the analysis of student essays. The aim of this research question is to understand whether XIP's analysis criteria can match those required by human markers for undergraduate essays in various disciplines. In doing this, quantitative analysis was conducted to explore how successful XIP analysis is in identifying the quality of student essays in terms of essay grades. The data collected for this research question therefore includes a range of argumentative student essays from different disciplines (hard

knowledge fields like physical sciences, and soft knowledge fields like arts and humanities, social sciences) with their essay grades. All these student essays were rendered through XIP and the results were quantified in terms of the number of salient sentences for each XIP category. The dataset was then analysed using the quantitative method of regression analysis, as described in detail in chapter 5.

RQ2: *How do educators define the attributes of good student writing and to what degree can the automated rhetorical parser XIP identify the presence of these attributes?*

The second sub-research question requires educators' understanding of what makes good student writing and to what degree this overlaps with how XIP judges good writing. RQ1 used a quantitative measure of essay grades as a way of identifying good quality student writing. For the RQ2, qualitative measure of the educators' definition of a good quality student writing was used. In order to understand and explore the opinions of educators on this matter, individual interviews were conducted with educators after they had assessed a student essay. In order to answer this question, it was necessary to observe educators in their real-world context of essay assessment and to comprehend their way of thinking about what makes a good essay; therefore, a qualitative method of data collection was chosen. As educators mostly assess student essays alone, the observation needed to be in an individual setting, which also allowed further questions to be asked. This facilitated the exploration of what educators think makes good student writing, what the attributes of good student writing are and what they value most when assessing student writing. Interview recordings were transcribed by using an intelligent verbatim approach[6] (McLellan, MacQueen, & Neidig, 2003). The emerging themes from the thematic analysis of interview data were used to evaluate to what degree XIP output can meet the educators' expectations. Following this, an overlap study between educators' and XIP's highlights

[6] Leaving out background noises, 'um', 'eh', 'you know' etc. for both interviewer and respondents but includes ((*laughs*)) to show emotion of respondent. No repetitions, descriptors or tripping over words included. The transcript was tidied up but without losing or adding to any of the important data.

was used to measure levels of accuracy and how well XIP's highlights aligned with the educators'. For a detailed description of the study, data collection and analysis see chapter chapter 6.

RQ3: *How congruent is the XIP analysis of student essays with educators' judgement of quality?*

The third sub-research question explores how reliable educators find the XIP output in terms of the quality of student essays, whether this type of analysis by XIP is acceptable to them, and whether XIP results are reliable and trustworthy enough to embed within a pedagogical context in higher education.

Focus group interviews with educators and senior researchers in the area of academic writing were carried out not only to add to the reliability of the quantitative data analysis results, but also to observe and comprehend the ideas and the interaction between experts about the applicability of XIP in an educational context, and to discuss possible actions to make it better for the purpose. Group interaction is the essential feature of the focus group method; it is distinguished from one-to-one interviews or questionnaires by the opportunities it provides to reflect on the interaction between research participants. The ability to observe participants' agreement and disagreement is an important strength of the focus group method. Therefore, in order to answer this research question, the focus group method was used. Focus group recordings were transcribed by using a full verbatim transcription (Poland, 1995). The qualitative method of thematic analysis was used for data analysis of the transcribed focus group session. In quantitative data analysis of any sort of transcribed talk, the analytic possibilities are reduced at an early stage and researchers work with pre-defined categories, and this would have limited the analyst's sensitivity to what was taking place in the talk. Instead, a qualitative approach was used, so that emerging themes and categories were generated based on the data rather than on prior

assumptions underlying the coding scheme used (Mercer, 2010). For a detailed description of the study, data collection and analysis see chapter 7.

RQ4: *In what ways should the output of the XIP analysis of student essays be delivered to educators so that they make use of this output to give feedback on student essays?*

The final sub-research question explores the ways in which XIP output can be presented to educators so that they make use of this output to give valuable feedback on student essays. This research question was answered through a pilot study followed by a main study. The pilot study was the first attempt to make sense of the raw XIP analysis output through a user-friendly dashboard.

In doing this, it was necessary to collect user requirements for the possible output. This was done through discussions about paper prototypes with PhD students at the Open University in the pilot study, and with educators (module team chairs who take part in the production of the module) in the main study. This elicited initial reactions from experts as well as their requirements. Co-designing a prototype tool with experts led to a working prototype. The aim was not to create a product but to use a working prototype which used a combination of previous findings to develop guidelines that may inspire future researchers to develop a fully working tool. Therefore, the prototype was not tested in terms of the user experience or usability. For a detailed description of the study, data collection and analysis see chapter chapter 8.

The table below summarises the mixed-methods approach used for each research question.

Table 4.1 Data collection and analysis methods used for each research question

Research question	Data collection	Data analysis methods
1. To what extent can the automated rhetorical parser XIP be used to identify indicators of good academic writing in undergraduate student essays from different disciplines, as judged by the essay grade?	• Qualitative data: student essays • Quantitative data: essay grades	• XIP analysis of student essays • Quantification of XIP analysis results • Regression analysis with quantified XIP

		analysis results and essay grade
2. How do educators define the attributes of good student writing and to what degree can the automated rhetorical parser XIP identify the presence of these attributes?	• Qualitative data from individual interviews with educators • Quantitative data of the number of sentences that educators highlighted	• Qualitative thematic analysis of the transcribed interviews • Quantitative measurement of the degree of overlap between XIP highlights and educators' highlights on student essays
3. How congruent is the XIP analysis of student essays with educators' judgement of quality?	Qualitative data: focus group session with educators	Thematic analysis of the transcribed focus group session
4. In what ways should the output of the XIP analysis of student essays be delivered to educators so that they make use of this output to give feedback on student essays?	Qualitative data: focus group session with educators to elicit the user and system requirements with educators	Thematic analysis of the transcribed focus group session

Overall the data collection methods described above were selected because they were the most suitable for answering the specific research question; data analysis methods were selected considering their suitability to the data as well as the research question.

4.5 Research design

The main study begins, in Chapter 5, with an exploration of how well the current build of XIP works for the analysis of student essays in different disciplines. This involved analysing various student essays with XIP across disciplines, and judging the results based on the awarded essay grade. For this study, quantitative regression analysis was used. Following this, in Chapter 6, the research examines how much of this analysis is acceptable to educators. It considers whether or not XIP results are reliable and trustworthy

enough for the tool to be embedded within a pedagogical context in higher education. This involves collecting educators' perspectives through focus group interviews.

Then, in Chapter 7 prior to the adoption of XIP technology into educational practices, the research investigates what educators value most in student writing, what the good features of student writing are, and what makes a good essay. This was done through individual interviews with educators who also assessed a sample of student writing in the session in which they were asked to highlight the key sentences they thought contributed positively to the essay grade. These sentences were then compared with XIP's salient sentences to measure the level of accuracy of XIP's output and to understand to what degree XIP's highlights overlap with those of the educators.

Finally, following an initial pilot study, a working prototype of XIP output was co-designed with educators, as described in Chapter 8. Possible ways of presenting the XIP output to educators who can make use of this output to give feedback on student essays were explored and then implemented. The prototype was evaluated to elicit educators' reactions, the aim of which was to create guidelines which can inspire future researchers to develop a fully working tool from this prototype.

Each of these studies is explained in the following chapters, and data collection, analysis and results are described in detail. Although these chapters also describe the reliability and validity aspects of the methods, the next section provides a brief overview of these matters.

4.6 Reliability, validity and generalisability

This section provides a succinct overview of how this thesis locates the concepts of reliability and validity in its studies. Since a mixed-methods approach is used in this research, it is necessary to discuss reliability and validity for both quantitative and qualitative studies.

Reliability and validity concepts were developed in the natural sciences, and there are various views and concerns about whether the same concepts have any value in

determining the quality of qualitative research (Jane Lewis & Ritchie, 2003). "Although the majority of today's qualitative researchers agree about the need for truthfulness in their research, they do not all agree with their quantitative counterparts about the need for reliability and validity as defined in quantitative research" (Wolcott, 2005).

In quantitative research, reliability refers to the replicability of research findings and whether or not the same findings would repeatedly come up if different studies were conducted, using the same or similar methods (J. Kirk & Miller, 1986). Validity refers to the correctness, truthfulness or precision of a research study. It determines whether the research truly measures that which it was intended to measure (Joppe, 2000).

Since reliability and validity are rooted in a positivist perspective, they should be redefined for their use for qualitative research (Golafshani, 2003). Constructivism, in qualitative research, values multiple realities that people have in their minds. "Therefore, to acquire valid and reliable multiple and diverse realities, multiple methods of searching or gathering data are in order" (Golafshani, 2003). Because of the nature of the qualitative research, researchers have come up with various terms for validity and reliability that have similar meanings. These include dependability (Lincoln & Guba, 1985), Clont, 1992 and Seale, 1999), trustworthiness (Seale, 1999), and confirmability and consistency (Lewis and Ritchie, 2003).

The reliability and validity concepts are important criteria for quality in quantitative research. However, reliability and validity are important factors which qualitative researchers should also be concerned about when designing a study, analysing results and judging the quality of the study (Patton, 2001). Although reliability and validity are two different concepts in quantitative research, in qualitative research these two are not usually treated separately. Instead, the same terminology is used to refer to both. Reliability in qualitative research is also synonymous with consistency. Qualitative research is considered reliable if the research findings can be replicated by another researcher as

Schwandt (2015) noted: “Traditionally, social scientists assume that while not all repeatable and replicable observations and accounts are necessarily valid, all valid accounts are (at least in principle) replicable” (p. 262). However, replicability of data in qualitative research may not always be possible. “Replicability of data may be difficult or impossible in some kinds of research: interviewees may die or disappear, and direct observations of real-world events by witnesses or participants cannot be repeated” (King, Keohane, & Verba, 1994, p. 26). Although researchers cannot travel back in time, and replicability cannot always be perfect, but it can provide valuable advances to the field. If researchers attempt to achieve as much replicability as possible by keeping complete notes and records about their methods, and the rules, data and procedures used to gather information and their inferences, another researcher can do the same thing and draw the same conclusion (King *et al.,* 1994). In qualitative research, “validity refers to the extent to which the phenomenon under study is being accurately reflected, as perceived by the study population” (Lewis and Ritchie, 2003).

In the quantitative sections of this thesis, reliability was established by repeating the XIP analyses across disciplines: various student essays from four different disciplines were analysed by XIP. Then, for each set of student essays in each discipline, reliability was maintained by repeating the regression analyses internally through repeated regression analysis of a randomly selected subset of these essays. Understanding the extent to which the results are consistent across different disciplines is a way of ensuring reliability, as the results of a study can be reproduced using a similar methodology, which means the research instrument is considered to be reliable. To determine validity, i.e. how truthful the results were, qualitative data analysis was carried out with educators to gain a better understanding of XIP’s performance.

As Lewis (2009) details, qualitative researchers have numerous methods available to them to enhance the evidence of reliability and validity. Adopting multiple methods, such as observation, interviews and recordings will lead to more valid, reliable and diverse

constructions of realities. Triangulation "involves the use of multiple and different methods, investigators, sources and theories to obtain corroborating evidence" (Onwuegbuzie & Leech, 2007, p. 239). It is a step taken by qualitative researchers to obtain several interpretations of the data at different times or in different locations by additional peer researchers (Johnson, 1997). According to Mathison (1988), triangulation is a test to improve the reliability and validity of the qualitative research and evaluation of findings. "Triangulation as used in quantitative research to test the reliability and validity can also illuminate some ways to test or maximize the validity and reliability of a qualitative study" (Golafshani, 2003). Triangulation can be carried out by using multiple methods while collecting data, such as observations and notes during interviews, and recordings to ensure reliability and validity. All the qualitative data collected in this research includes audio-recordings along with notes and observations. Additionally, triangulation is used throughout in the interpretation of the qualitative data. To achieve validity and reliability in the interpretation of the data, two peer researchers coded the interview data independently to eliminate bias, and the researcher repeated the qualitative thematic analysis and coding at different times and in different locations.

The next chapter is the first analysis chapter, a quantitative study that addresses the first subsidiary research question 'To what extent can the automated rhetorical parser XIP be used to identify indicators of good academic writing in undergraduate student essays from different disciplines, as judged by the essay grade?'.

[This page intentionally left blank.]

5

THE XIP ANALYSIS OF STUDENT WRITING

5.1 Introduction

This chapter addresses the first subsidiary research question: "To what extent can the automated rhetorical parser XIP be used to identify indicators of good academic writing in undergraduate student essays from different disciplines, as judged by the essay grade?" Answering this question required sets of undergraduate student essays and their essay marks to be analysed. As the literature suggested that student writing differs by subject area and level, various datasets from different disciplines and levels needed to be analysed through XIP, in order to explore how well the XIP works across disciplines.

As previously discussed, there is a need to validate XIP in higher education contexts, since it was originally developed in the context of academic literature for the analysis of journal papers. This chapter describes an evaluation study of XIP on undergraduate student essays from various disciplines (hard and soft disciplines as described in table 2.6sss) and levels (undergraduate years from one to three), using the mark awarded as a measure of the quality of the writing. As part of this exploration, the studies presented in this chapter (which are explained in more detail in the following sections) were designed to assess the quality and generalisability of XIP through correlational studies and regression analysis.

The studies conducted for answering the first subsidiary research question are presented in four sections, each dealing with one of the datasets used for the studies. The British Academic Written English (BAWE) corpus, and three Open University modules named here as S000, E000 and L000 (module codes) were the datasets used for the subsequent analysis. Each section gives the description of the dataset, explains the nature of the studies carried out, reports the results and concludes with a discussion of these results. The following section describes the overall nature of the studies carried out which were designed to answer the first subsidiary research question.

5.2 Study

Although several datasets were used, the process of the studies remained the same. Once the student essays were collected in .txt or .doc format, the following steps were repeated with each dataset through an infrastructure[7] developed by the researcher for automating the process which:

1. Parses all the essays through the infrastructure for converting each student essay into .xml format by adding special XML headers required for XIP analysis in order to clean the data so it can be run through XIP.
2. Analyses each .xml formatted student essay with XIP.
3. Counts the number of salient sentences that XIP analysis produced (the total number of salient sentences in each essay and total number of each sentence categories/types per essay) for quantitative statistical analysis.
4. Exports these numerical results into an Excel sheet to conduct quantitative statistical functions on them.

[7] Two minutes introductory YouTube video about the infrastructure: https://www.youtube.com/watch?v=KJPvgrX_Lbg

The following diagram summarises the infrastructure implemented for this study.

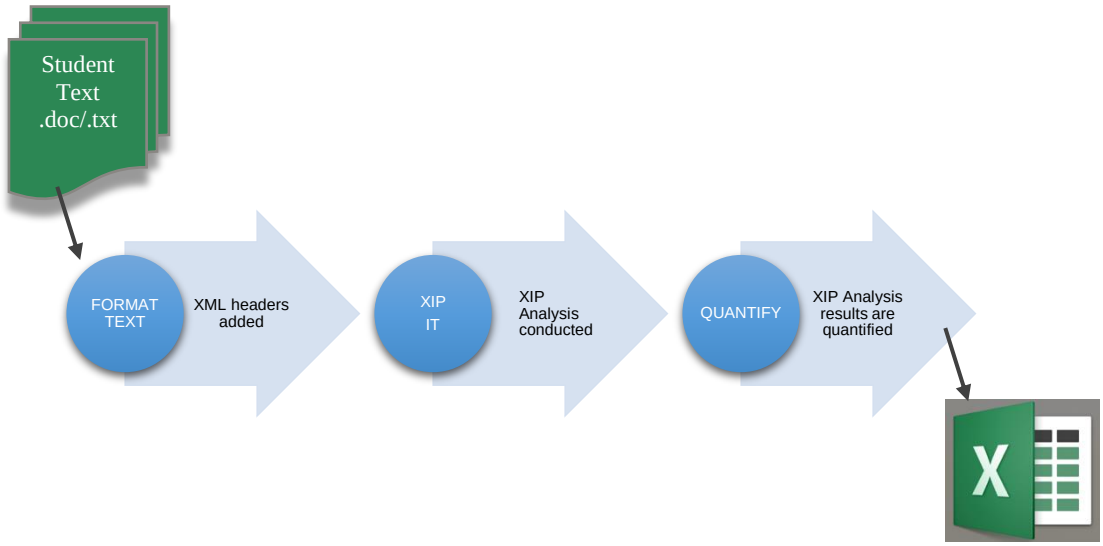


Figure 5.1 Infrastructure implemented for the quantitative studies of this thesis

Once the results were exported through the infrastructure, depending on the form of the dataset (i.e. variable types), the correlational and/or regression analyses were conducted through IBM's SPSS statistical software package.

For the correlational studies, the frequency of XIP categories were correlated with the essay mark in order to understand whether XIP categories have some relation with the given mark, which is used as an indication of the quality. Correlation measures the strengths of association between two variables, and in statistics there are three types of correlation: Pearson correlation, Kendall rank correlation and Spearman correlation. Pearson's, Spearman's and Kendall's correlation coefficients are the most commonly-used measures but the latter two are usually recommended for non-normally distributed data (Bobko, 2001). Therefore, for this study the Pearson correlation coefficient was used when measuring the degree of relationship between linear related variables of the essay mark and numbers of sentences in each XIP category.

The strength of the correlation coefficient results were interpreted based on Dancey and Reidy (2004) categories; which are widely used for such interpretations (Table 5.1):

Table 5.1 Dancey & Reidy's correlation categories used for interpretation of results

Value of the Correlation Coefficient	Strength of Correlation
1	Perfect
0.7 - 0.9	Strong
0.4 - 0.6	Moderate
0.1 - 0.3	Weak
0	No meaningful correlation

Correlational analyses enabled exploration of whether XIP is capable of identifying good indicators of student writing. Regression analyses were then carried out following the correlational studies, where possible (depending on the type of the variables within the dataset). "Regression analysis is a conceptually simple method for investigating functional relationships among variables" (Chatterjee & Hadi, 2012, p. 1). Regression analysis provides an index describing the linear relationship between two variables (total number of salient sentences and the essay mark) in order to predict the relationship between more than two variables to identify which of these variables (XIP categories), if any, can predict the outcome variable (essay mark) (Crawford, 2006). Regression analysis takes the correlational studies further in order to explore which XIP categories (individually and together) have an effect on the essay mark. Regression analyses help to interpret which XIP categories affect a given mark, and to what degree. In these analyses, the models were specified in such a way that the dependent variable was a linear combination of independent variables. Therefore, linear regression models were used for this study and non-linear regression analysis alternatives were rejected. In the linear regression models, the mark of the essays was taken as the dependent variable and the total numbers of salient sentences for each XIP category (TENDENCY, EMPHASIS, NOVELTY, SUMMARY, OPEN QUESTION, CONTRAST and BACKGROUND, see chapter 3.8.3 for the definitions), marked up in the essays were taken as independent variables. In other words, these studies regressed the assigned mark on each XIP category, which was computed by IBM's

SPSS statistical software package (Field, 2013) using the linear regression functions. Following normal convention, interpretation of the regression analyses is made as follows: $p \leq 0.05$ signifies a statistically significant result, and $p \leq 0.01$ is regarded as highly significant; and the rest is not significant.

In SPSS, there are two modified versions of R-square, one developed by Cox and Snell (1989) and the other developed by Nagelkerke (1991). Since the former R-square value cannot reach 1, the latter modified it. The correction increases Cox and Snell's version to make 1 a possible value for R-square. Therefore, to evaluate the power of explanation of the logistic regression model, Nagelkerke's R-square is used. Nagelkerke's R-square will always be less than the Cox and Snell R-square, since the Nagelkerke's R-square is an adjustment of Cox and Snell, for which the maximum value it can attain is equal to 1 (the maximum value for Cox and Snell is 0.75). Although it looks like simply reporting the larger value, the case here is simply a matter of scaling, not actual size difference.

Additionally, in SPSS there are two types of coefficients that are displayed in a multiple regression tables: unstandardised coefficients, and standardised coefficients. Standardised regression coefficients are based on changes in standard deviation units; but unstandardised regression coefficients are based on metric changes. To illustrate this, with unstandardised regression every metric unit change in the independent variable changes the dependent variable by X units. One advantage of using unstandardised coefficients is that they provide readily interpretable substantive meaning. The major problem with standardised coefficients is that they lack the property of causal invariance, but unstandardised regression coefficients do reflect casual invariance. Thus, unstandardised coefficients are generally preferred to standardised ones (Jaccard, Wan, & Turrisi, 1990). Unstandardised coefficients are generally used when the raw units are familiar, such as years, kilograms, or centimetres. However, if the variables' raw units are not well-known in everyday usage, then standardised coefficients are used. Since the values withint the

BAWE corpus are ‘grade’ and ‘number of sentences’ which are familiar everyday terms, the unstandardised coefficient is used for the interpretations given in following sections. The next section describes the first dataset used for the XIP’s performance validation study.

5.3 BAWE DATASET

5.3.1 Background

The research project *An Investigation of Genres of Assessed Writing in British Higher Education*[8], was undertaken between 2004 and 2007 at the universities of Warwick, Reading and Oxford Brookes. The British Academic Written English (BAWE) corpus (Nesi, 2011), the result of this project, is a collection of good quality student writing which were marked as either gaining merit or distinction.

Student writing for the BAWE project was collected from all three participating universities. Texts were collected from 35 disciplines in four broad disciplinary groupings, and from students in each of three undergraduate years (referred to in this thesis as ‘levels’, since students at The Open University who take first level modules are not necessarily in their first year of university study) and those on masters’ courses. For the study of this thesis, the undergraduate levels were taken into account. The total number of student participants is 1,039 from whom the undergraduate levels were taken into account for the study of this thesis. There are three levels of undergraduate years and the numbers of students from each is 333, 302, and 234 respectively.

In the study of this thesis, broad disciplinary groupings were used when making judgements on the performance of XIP because under each specific discipline the numbers of assignments (as shown in black in Table 5.2) submitted were not significant enough to

[8] The BAWE Corpus Manual: http://goo.gl/D4mQrd

make valuable statistical interpretations (for valuable interpretations the sample size needs to be at least 100, preferably 200 (King & Zeng, 2001)). The smaller the sample size, the greater the likelihood of obtaining less rigorous and reliable interpretations. Therefore, this study takes four broad groupings into account; each of these contains more than 200 assignments. Below, in Table 5.2, red text shows the number of assignments submitted for each broad disciplinary grouping and green text shows the total number of students who submitted assignments at each level.

Table 5.2 Number of assignments in the BAWE corpus by discipline & level

Disciplinary group	Discipline	Lvl 1	Lvl 2	Lvl 3	Masters	Total
Social Sciences	Anthropology	14	12	6	17	49
	Business	32	33	31	50	146
	Economics	30	30	23	13	96
	Hospitality, Leisure and Tourism Management (HLTM)	14	21	29	29	93
	Law	37	37	31	28	134
	Politics	37	33	15	25	110
	Publishing	11	4	0	15	30
	Sociology	32	25	24	21	110
	Other	0	2	3	4	9
*Total number of **students**:*		*85*	*88*	*75*	*62*	***313***
*Total number of **assignments**:*		*207*	*197*	*162*	*202*	***777***
Arts & Humanities	Archaeology	23	21	15	17	76
	Classics	33	27	15	7	82
	Comparative American Studies	29	26	13	6	74
	English	35	35	28	8	106
	History	30	32	31	3	96

	Linguistics	27	31	24	33	115
	Philosophy	43	34	25	4	106
	Other	19	22	9	0	50
*Total number of **students**:*		*101*	*83*	*61*	*23*	***268***
*Total number of **assignments**:*		*239*	*228*	*160*	*78*	***705***
Life Sciences	Agriculture	35	35	30	34	134
	Biological Sciences	52	50	26	41	169
	Food Sciences	26	36	32	30	124
	Health	35	33	12	1	81
	Medicine	0	0	0	80	80
	Psychology	32	39	13	11	95
*Total number of **students**:*		*74*	*71*	*42*	*46*	***233***
*Total number of **assignments**:*		*180*	*193*	*113*	*197*	***683***
Physical Sciences	Architecture	2	4	2	1	9
	Chemistry	23	24	29	13	89
	Computer Science	34	13	30	10	87
	Cybernetics & Electronics	4	4	13	7	28
	Engineering	59	71	54	54	238
	Mathematics	8	5	12	8	33
	Meteorology	6	9	0	14	29
	Physics	37	14	14	3	68
	Planning	8	4	2	0	14
	Other	0	1	0	0	1
*Total number of **students**:*		*73*	*60*	*56*	*36*	***225***
*Total number of **assignments**:*		*181*	*149*	*156*	*110*	***596***
OVERALL STUDENT NUMBER		**333**	**302**	**234**	**167**	**1039**
OVERALL ASSIGNMENT NUMBER		**807**	**767**	**591**	**596**	**2761**

5.3.1.1 Genre family

In order to make sure that the overall dataset allows for reasonable statistical interpretations, the assignment type was carefully selected considering the total number of assignments. All student assignments in the corpus were scrutinised for generic properties, and a large number of genres were identified by the BAWE project members. These genres were collected into thirteen genre families, classes of genres sharing functional and structural properties. For the study of this thesis, genre families containing argumentative writing were selected only if the descriptions given (see APPENDIX A) and the student writing for each of those tallied well with the intended use for XIP, and with what XIP is intended to extract from writing. This front selection process revealed that XIP is not suitable for all genres. After the filtering process, five of the 13 genre families were selected: *essay*, *literature review*, *problem question*, *proposal*, and *research plan*. The descriptions of each selected genre family are given below:

- Essay: to develop the ability to construct a '*coherent argument'* and develop '*critical thinking skills';* these may be discussion (issue, pros/cons, final position), exposition (thesis, evidence, restate thesis), factorial (outcome, conditioning factors), challenge (opposition to existing theory), comparison (series of comparative points or arguments), or commentary (series of comments on a text); they may correspond to a published academic/specialist paper
- Literature Review: to demonstrate familiarity with literature relevant to the focus of study; includes summary of literature relevant to the focus of study and varying degrees of '*critical evaluation'*
- Problem Question: to practise applying specific methods in response to simulated professional problems; problem (may not be stated in assignment), '*application of relevant arguments'* or presentation of possible solution(s) in response to scenario

- Proposal: to demonstrate ability to make a case for future action; includes purpose, detailed plan, '*persuasive argumentation'*
- Research Plan: to demonstrate ability to undertake a complete piece of research including research design, and an appreciation of its significance in the field; may include Literature Review, Methods, Findings, Discussion.

The table below overviews the total number of assignments used in this study from each disciplinary grouping after filtering based on the genre family. (Note that the dataset for the *Physical Sciences* has fewer than 100 assignments, the possible effect of which is discussed later in the chapter.)

Table 5.3 The total number of assignments used for the BAWE study

Disciplinary grouping	Number of assignments analysed
Arts & Humanities	541
Life Sciences	141
Physical Sciences	80
Social Sciences	352

5.3.2 BAWE study

The first subsidiary research question explores how well the current build of XIP works for the analysis of student essays in different disciplines which involves analysing various student essays with XIP across disciplines, and judging the results based on the essay grade awarded. With this in mind, the BAWE dataset was selected for the following reasons:

1. The BAWE corpus consists of four different disciplinary groupings: arts and humanities, life sciences, physical sciences and social sciences. This supported reflection on the argument, as discussed in literature review, that the nature of student writing is discipline-specific. The corpus also permitted exploration of which subject area is most suited to be supported with automated rhetorical parser, XIP.
2. The BAWE corpus consists of student writing from different levels (from undergraduate year one to three, and masters level). As the literature review

showed that student writing also differs by level, this dataset enabled reflection on how XIP performs at different student levels as well as in subject domains.

3. The BAWE corpus consists of student writing from different levels and domains that were marked either as merit or distinction. As the research question investigates to what extent automated rhetorical parser XIP can identify '**good'** indicators of student writing, the highly marked essays contained in this dataset allowed an understanding of what specifically the features of good essays are, and what emerges when they are analysed by XIP.

Unlike in the other datasets analysed, the numeric essay marks were not available in the BAWE corpus. Within the corpus, the mark is provided either as distinction or merit. Since the mark is a dichotomous variable, correlation studies could not be conducted. Due to the dichotomous variable mark, *binary logistic linear regression analysis* was carried out with the BAWE dataset because in situations where the dependent variable is dichotomous (0 for merit, 1 for distinction), the most common procedure is to use *binary logistic regression* (Tranmer & Elliot, 2008).

5.3.3 Results

Binary linear logistic regression models for each disciplinary grouping and each level under these groupings were tested separately in order to gain a better understanding of how XIP performs in different disciplines and levels. The following table provides an overview of the results. These are described in more detail in the following sections.

Table 5.4 The regression analysis results for the BAWE corpus, showing which XIP categories influence the essay grade across levels and disciplines

Disciplinary group	Level 1	Level 2	Level 3
Social Sciences	NOT SIGNIFICANT	SUMMARY (p=0.007)	CONTRAST (p=0.039) NOVELTY (p=0.02)

Arts and Humanities	EMPHASIS (p=0.033)	SUMMARY (p=0.005)	NOT SIGNIFICANT
Life Sciences	NOT SIGNIFICANT	TENDENCY (p≤0.001)	OPEN QUESTION (p=0.002)
Physical Sciences	NOT SIGNIFICANT	NOT SIGNIFICANT	NOT SIGNIFICANT

5.3.3.1 Social sciences

In total, the Social Sciences discipline has 352 texts for all three undergraduate levels. The initial regression model was tested with **all levels**. The XIP categories were added to the model as independent variables which can affect the dependent variable essay mark. The regression model did **not** prove to be **significant**. To understand XIP's behaviour at each level, individually, three logistic regression models were tested for every level. Summary of the results are given in the following table, which are then described in detail in following sections.

Table 5.4.1 The regression analysis results for the social sciences, showing the significance of the models, and which XIP categories influence the essay grade

Level	Number of essays analysed		Significance of the regression model	Nagelkerke's R-Square	Significant XIP Categories
Level 1	139		NOT SIGNIFICANT (p=0.456)	N/A	NONE (p>0.05)
	87 merits	52 distinctions			
Level 2	130		HIGHLY SIGNIFICANT (p≤0.01)	0.094	SUMMARY (p=0.007)
	86 merits	44 distinctions			
Level 3	83		HIGHLY SIGNIFICANT (p≤0.01)	0.171	CONTRAST (p=0.039) NOVELTY (p=0.02)
	60 merits	23 distinctions			

Level 1 -SS

One hundred and thirty-nine essays (87 merits and 52 distinctions) were used for testing the model. The regression model did **not** prove to be **significant** (p=0.456) and none of the XIP categories (p>0.05 for each) were found to be statistically significant for Social Sciences, Level 1.

Level 2 -SS

One hundred and thirty essays (86 merits and 44 distinctions) were used for testing the model. The regression model proved to be highly significant ($p \leq 0.01$).

Based on the Nagelkerke's R-square:

- the statistical results indicated that 9.4% of the variation in essay mark is explained by the model which means that independent variables in the model affect the mark by 17.1% (Nagelkerke's R-Square=0.094).

When each independent variable was analysed, it was found that the category SUMMARY was highly significant and had explanatory power for the dependent variable essay mark (SUMMARY, p=0.007). When unstandardised coefficients were examined for SUMMARY, the following interpretation could be made:

- *In this dataset, when the number of SUMMARY sentences increases by one unit, the odds that the student is awarded a distinction increase by a factor of 1.222, when other variables are controlled.*

Level 3 -SS

Eighty-three essays (60 merits and 23 distinctions) were used to test the model. The regression model proved to be highly significant ($p \leq 0.01$).

Based on Nagelkerke's R-square:

- the statistical results indicated that 17.1% of the variation in essay mark is explained by the model which means that independent variables in the model affect the mark by 17.1% (Nagelkerke's R-Square=0.171).

When each independent variable was analysed, it was found that the categories CONTRAST and NOVELTY were statistically significant and had an explanatory power for the dependent variable essay mark (CONTRAST p=0.039, NOVELTY p=0.02). When unstandardised coefficients were examined for CONTRAST and NOVELTY, the following interpretations can be made:

- *When the number of CONTRAST sentences increases by one unit, the odds that the student is awarded a distinction increase by a factor of 0.844, when other variables are controlled.*
- *When the number of NOVELTY sentences increases by one unit, the odds that the student is awarded a distinction increase by a factor of 1.759, when other variables are controlled.*

5.3.3.2 Arts and humanities

In total, the Arts and Humanities section of the BAWE corpus includes 541 essays. The initial regression model was tested with all levels. The XIP categories were added into the model as independent variables which can affect the dependent variable essay mark. It was found that the category SUMMARY ($p \leq 0.001$) had an explanatory power for the dependent variable, meaning it has a highly significant effect on the mark compared to other XIP categories.

- *For the Arts and Humanities discipline in general, when the number of SUMMARY sentences increases by one unit, the odds that the student gets a distinction increase by a factor of 1.093, when other variables are controlled.*

To understand XIP's behaviour on each level, three logistic regression models were tested for every level whose results are summarised below (table 5.4.2), which are then described in following sections.

Table 5.4.2 The regression analysis results for the arts and humanities, showing the significance of the models, and which XIP categories influence the essay grade

Level	Number of essays analysed		Significance of the regression model	Nagelkerke's R-Square	Significant XIP Categories
Level 1	215		HIGHLY SIGNIFICANT ($p \leq 0.01$)	0.047	EMPHASIS ($p=0.033$)
	159 merits	56 distinctions			
Level 2	194		HIGHLY SIGNIFICANT ($p \leq 0.01$)	0.059	SUMMARY ($p=0.005$)
	143 merits	51 distinctions			
Level 3	132		NOT	N/A	NONE

	73 merits	59 distinctions	SIGNIFICANT (p=0.972)		(p>0.05)

Level 1 -AH

Two hundred and fifteen essays (159 merits and 56 distinctions) were used to test the model. The regression model proved to be highly significant ($p \leq 0.01$).

Based on the Nagelkerke's R-square:

- the statistical results indicated that 4.7% of the variation in essay mark is explained by the model which means that independent variables in the model affect the mark by 4.7% (Nagelkerke's R-Square=0.047).

When each independent variable was analysed, it was found that the category EMPHASIS was statistically significant and had an explanatory power for the dependent variable essay mark (EMPHASIS, p=0.033). When unstandardised coefficients were examined, the following interpretation could be made for the category EMPHASIS:

- *When the number of EMPHASIS sentences increases by one unit, the odds that the student is awarded a distinction also increase by a factor of 1.173, when other variables are controlled.*

Level 2 -AH

One hundred and ninety-four essays (143 merits and 51 distinctions) were used to test the model. The regression model proved to be highly significant ($p \leq 0.01$).

Based on Nagelkerke's R-square:

- the statistical results indicated that 5.9% of the variation in mark was explained by the model which means that independent variables in the model affect the mark by 5.9% (Nagelkerke's R-Square=0.059).

When each independent variable was analysed, it was found that the category SUMMARY was statistically significant and had an explanatory power for the dependent variable essay

mark (SUMMARY, p=0.005). When unstandardised coefficients were examined, the following interpretation could be made for the category SUMMARY:

- *When the number of SUMMARY sentences increases by one unit, the odds that the student is awarded a distinction also increase by a factor of 1.148, when other variables are controlled.*

Level 3 -AH

One hundred and thirty-two essays (73 merits and 59 distinctions) were used for testing the model. The regression model proved not to be significant (p=0.972). When coefficients were examined, none of the XIP categories was found statistically significant for Arts and Humanities, Level 3.

5.3.3.3 Life sciences

In total, the Life Sciences discipline of the BAWE corpus includes 141 essays for all three levels. When all levels were included in the regression model, it was found that category SUMMARY had a significant effect on the mark (SUMMARY, p≤0.017).

- *For the Life Sciences discipline in general, when the number of SUMMARY sentences increases by one unit, the odds that the student is awarded a distinction also increase by a factor of 1.139, when other variables are controlled.*

To test XIP's performance on each level, three logistic regression models were tested for every level whose results are summarised below (table 5.4.3), which are then described in following sections.

Table 5.4.3 The regression analysis results for the life sciences, showing the significance of the models, and which XIP categories influence the essay grade

Level	Number of essays analysed		Significance of the regression model	Nagelkerke's R-Square	Significant XIP Categories
Level 1	50		NOT SIGNIFICANT (p=0. 611)	N/A	NONE (p>0.05)
	29 merits	21 distinctions			
Level 2	63		SIGNIFICANT	0.035	TENDENCY

	34 merits	29 distinctions	(p=0.007)		(p≤0.001)
Level 3	28		SIGNIFICANT (p=0.013)	0.0625	OPEN QUESTION (p=0.002)
	14 merits	14 distinctions			

Level 1 -LS

Fifty essays (29 merits and 21 distinctions) were used for testing the model. The regression model proved not to be significant (p=0.611). When coefficients were examined, none of the XIP categories was found statistically significant for Life Sciences, Level 1.

Level 2 -LS

Sixty-three essays (34 merits and 29 distinctions) were used for testing the model. The model proved to be significant (p=0.007).

Based on the Nagelkerke's R-square:

- the statistical results indicated that 3.5% of the variation in essay mark is explained by the model which means that independent variables in the model affect the mark by 3.5% (Nagelkerke's R-Square=0.035).

When each independent variable was analysed, it was found that the category TENDENCY (sentences describing research trends and emerging research directions) was statistically significant and had an explanatory power for the dependent variable essay mark (TENDENCY, p≤0.001). When unstandardised coefficients were examined, the following interpretation could be made for the category TENDENCY:

- *When the number of TENDENCY sentences increases by one unit, the odds that the student is awarded a distinction also increase by a factor of 14.459, when other variables are controlled.*

Level 3 -LS

Twenty-eight essays (14 merits and 14 distinctions) were used for testing the model. The model proved to be significant (p=0.013).

Based on the Nagelkerke's R-square:

- the statistical results indicated that 6.25% of the variation in essay mark is explained by the model which means that independent variables in the model affect the mark by 6.25% (Nagelkerke's R-Square=0.0625).

When each independent variable was analysed, it was found that the category OPEN QUESTION was statistically significant and had an explanatory power for the dependent variable essay mark (OPEN QUESTION, p=0.002). When unstandardised coefficients were examined, the following interpretation could be made for the category OPEN QUESTION:

- *When the number of OPEN QUESTION sentences increases by one unit, the odds that the student is awarded a distinction also increase by a factor of 6.923, when other variables are controlled.*

5.3.3.4 Physical sciences

In total, the Physical Sciences discipline of the BAWE corpus includes 80 essays for all three levels (37 merits and 43 distinctions). When all levels were included in the regression model, it was found that category NOVELTY had a significant effect on the mark (NOVELTY, p=0.03).

- *For the Physical Sciences discipline in general, when the number of NOVELTY sentences increases by one unit, the odds that the student is awarded a distinction also increase by a factor of 2.272, when other variables are controlled.*

To test XIP's performance on each level, three logistic regression models were tested for every level; however, none of the XIP categories was significant enough to make any level-based interpretations, as also shown in table 5.4.4.

Table 5.4.4 The regression analysis results for the physical sciences, showing the significance of the models, and which XIP categories influence the essay grade

Level	Number of essays analysed	Significance of the regression model	Nagelkerke's R-Square	Significant XIP Categories
Level 1	29	NOT	N/A	NONE

	12 merits	17 distinctions	SIGNIFICANT (p=0. 356)		(p>0.05)
Level 2	18		NOT SIGNIFICANT (p=0. 350)	N/A	NONE (p>0.05)
	11 merits	7 distinctions			
Level 3	33		NOT SIGNIFICANT (p=0. 386)	N/A	NONE (p>0.05)
	14 merits	19 distinctions			

5.3.4 Discussion of the BAWE analysis

After conducting binary logistic regression analyses with student texts from the BAWE corpus based on disciplinary groupings and levels under each group, it was noted that XIP performed differently for each discipline and each level. Table 5.5 summarises the statistically significant XIP categories for each discipline and level, including the p values and odds ratios.

Table 5.5 Summary of the binary logistic regression analysis results for BAWE corpus

Disciplinary group	Overall	Level 1	Level 2	Level 3
Social Sciences	NOT SIGNIFICANT	NOT SIGNIFICANT	SUMMARY (p=0.007, odds=1.222)	CONTRAST (p=0.039, odds=0.844) NOVELTY (p=0.02, odds=1.759)
Arts and Humanities	SUMMARY (p≤0.001, odds=1.093)	EMPHASIS (p=0.033, odds=1.173)	SUMMARY (p=0.005, odds=1.148)	NOT SIGNIFICANT
Life Sciences	SUMMARY (p=0.007, odds=1.139)	NOT SIGNIFICANT	TENDENCY (p≤0.001, odds=14.459)	OPEN QUESTION (p=0.002, odds=6.923)
Physical Sciences	NOVELTY (p=0.03, odds=2.272)	NOT SIGNIFICANT	NOT SIGNIFICANT	NOT SIGNIFICANT

The following table provides some examples of the XIP analysis, demonstrating the salient sentences identified by XIP for significant XIP categories; to show the accuracy of its results. For each statistically significant XIP category, two distinct examples are given.

The linguistic features, meta-discourse, shown in bold demonstrate the reason why XIP assigned this particular sentence category.

Table 5.6 Examples of the salient sentences picked up by the XIP from the BAWE corpus

	Level 1	Level 2	Level 3
SS		**SUMMARY** "*The aim of this essay is to expose the importance of China through a theoretical approach which places China within a larger capitalist world order."* *"The final section of this essay will discuss the possible disruption of such a chain with reference to China's domestic politics and the strain its role in the chain puts on domestic politics, thereby exposing the potential importance of China to the global order."*	**CONTRAST** "*All these overriding assumptions become problematic when we examine contemporary examples of victimization, when women are also agents of violence while men become the victims of sexual abuse.*" "*While women may play an assertive role in ethnic conflicts, they may not be military combatants fighting for a worthy cause; rather they could be active perpetrators of inhumane war crimes."* **NOVELTY** "*In a similar light, the ethnic conflicts that have engulfed the regions of the former Yugoslavia during the past decade offer us new perspectives of looking at the roles women play."* *"More importantly, its implications for gender studies are paramount; it will provide new insight for feminist accounts of women's agency in conflicts."*
AH	**EMPHASIS** *"Dante was an important figure in the development of Renaissance literature."* *"Florence had a key role in the development of the Italian Renaissance because its cultural pre - eminence coincided with the largest territorial expansion of the time ."*	**SUMMARY** *"Patterns of births, marriages, deaths and migrations helped to shape society and economy and were themselves shaped by society and economy, as I have attempted to show in this essay."* *"After all, the authors are aware that methodological shortcomings exist; this work does not pretend to be concrete evidence but merely a likely estimation."*	
LS		**TENDENCY** *"The use of verbal reports and discourse*	**OPEN QUESTION** *"Many foods contain non-nutritive components such as anthocyanins and other*

		analysis has increased in popularity as a research method and has resulted in many contributions to the study of psychology." *"Hopefully as research into discourse analysis increases, journals might adapt their guideline in order to cater more for qualitative research (Coyle, 1995)."*	*phenolic compounds which provide protection against chronic diseases through multiple effects are as yet poorly understood."* *"Very little is known about the absorption and metabolism of anthocyanins and whether they are absorbed in sufficient quantities and in a form in which effects on in vivo measures of oxidative cellular damage."*
PS			

The results demonstrated that XIP did not perform well on BAWE essays drawn from the Physical Sciences. The XIP categories do not work well for undergraduate science essays. Writing for hard knowledge disciplines requires different discourse (Kelly, 2007), which could explain this result. On the other hand, XIP's performance showed promising results on the other disciplines.

The highest odds ratios were found in the Life Sciences, meaning the XIP categories TENDENCY for Level 2 and OPEN QUESTION for Level 3 were highly effective on the given essay mark. Each of these sentence types increased the students' final essay mark significantly with odds of 14.459, and 6.923 respectively. These two categories did not come up in any other discipline or level. Considering that XIP was originally implemented to analyse the abstracts of journal papers from the life sciences disciplines, this result suggests that the XIP also works for the student writing in this disciplinary area. These two categories require higher-order writing skills; which are expected to be seen in the discourse moves of the experienced researchers' writing. As explained in the literature review, all of the XIP categories are created based on the linguistic analysis of the experienced researchers' writing; however not all categories require higher-order writing skills such as SUMMARY where the writer should summarise the goals or the results of

the article, or BACKGROUND where the writer needs to describe the previous knowledge, the literature. On the other hand, TENDENCY describes research trends and emerging research directions and OPEN QUESTIONS describes problems which have not been solved. In order for an undergraduate student to identify and then describe an unresolved problem requires analytical and critical thinking which is a higher-order writing skill. This is possibly the reason why this category came up in Level 3.

Similarly, such improvement in the ability to use higher-order categories was observed in the Social Sciences. While the category SUMMARY was effective for Level 2, CONTRAST and NOVELTY were effective on the students' essay mark for Level 3. 'Experienced' students needed to describe tensions, contrasts between ideas, models or research directions and describe new research ideas to get higher marks.

With the exception of the Arts and Humanities discipline, the XIP categories did not influence the essay mark for Level 1 student texts. This result supports the arguments given in the literature review (Sommers & Saltz, 2004; Wingate, 2012) that newcomers to the university struggle with writing, and especially with producing essays rich in argumentation, but develop this skill later in their studies.

However, the dataset has some drawbacks for making further interpretations. First of all, there is an inconsistency between the numbers of essays under each disciplinary grouping. Arts and Humanities and Social Sciences are relatively good datasets as they include over 200 essays, which allows statistically reliable interpretations; but Life Sciences and especially Physical Sciences are underrepresented disciplines in terms of the sample size. Especially for ensuring the accuracy of the 'binary' logistic regression analysis, the sample size needs to be at least over 100 and preferably 200 because of the small-sample bias problem. Logit coefficients are biased in small samples (under about 200) (King and Zeng, 2001). Although each disciplinary grouping has over 200 essays, when level-based considerations are made the numbers of texts decrease dramatically for Life Sciences (141) and Physical Sciences (80). The sample size especially for the Physical Sciences is

illustrative, but not definitive as it is too smalland therefore does not permit generalisations, which requires further investigation.

Additionally, the interpretations are based on disciplinary groupings and each include several sub-disciplines. Essays from each of these sub-disciplines also come from various different assignments from various institutions. The texts do not come from a single assignment and therefore do not carry similar features to be interpreted as a whole. Moreover, in all of these analyses, there is no measure to ensure that length of text is not predictive. Although statistically it would not simply mean that higher grade texts are longer, since all of these texts are highly graded either as merit or distinction, it could be the case that longer texts might contain more XIP sentences. Due to the scope of this dataset, this is discounted in the BAWE corpus, but texts are the same length in subsequent databases to overcome this possible issue. Finally, marking rubrics and learning outcomes are not available therefore instrumentation biases cannot be eliminated. Information regarding who marked these essays, using which marking guidelines, is unavailable. In short, in the BAWE corpus there are multiple assignments, genres of writing, institutions all combined, with no access to the grading criteria, or even the whole assignment task itself. Although the selection of assignment types was thoughtfully processed based on the given genre description in the BAWE manual, there is still a possibility that some of the texts written for an assignment were not graded much on argumentative writing.

The BAWE study results showed to what extent XIP can be used to identify good indicators of academic writing when level and domain based considerations are made. The BAWE study results showed which other datasets, such as soft disciplines (i.e. Arts and Humanities, Social Sciences and Life Sciences), can be included in the study in order to make the work repeatable, reliable, and generalisable. Understanding the extent to which the results are consistent across different disciplines is a way of ensuring reliability, as the results of a study can be reproduced using a similar methodology, which means the

research instrument is considered to be reliable. The drawbacks mentioned above create a challenge in terms of the validity and reliability of the study. Therefore, the extent to which XIP can be used to identify good indicators of academic writing cannot be discussed with confidence on the basis of this dataset. However, these results strongly suggest that XIP performs differently in each discipline as expected and therefore further studies are required with various other datasets for validity. The following three sub-sections describe these studies.

Based on the examination of the BAWE results, Physical Sciences were chosen for further analysis because it was the discipline least likely to award marks for rhetoric. This would support further investigation of XIP's performance on hard disciplines, to find out why it did not give significant results in the BAWE study, why XIP did not work for hard knowledge disciplines. Additionally, to validate the BAWE results, soft knowledge disciplines were chosen for further analysis as they were the disciplines most likely to award marks. The studies were designed in such a way that similar datasets were analysed (e.g. two different arts and humanities datasets) so that the findings could be validated.

5.4 S000 DATASET

5.4.1 Introduction

The second set of student writing came from one of the second-level undergraduate science modules of the UK-based distance education university, The Open University. The S000 course on practical science was launched in 2012 and caters for students from five different science strands: 1) chemistry and analysis, 2) environmental science, 3) earth and environment, 4) biology and health, and 5) physics and astronomy. For this study, students' written texts submitted for the end-of-module assessment (EMA, referred to in this thesis as an 'assignment') were used dating from 2012 and 2013. In the first year of this module, in 2012, 275 students completed the assignment and in the following year, in

2013, 564 students completed the assignment. This dataset was selected for the following reasons:

1. As the BAWE corpus results demonstrated that XIP does not perform well on hard disciplines, there was a need to investigate this further as the reason could be that the essays in the BAWE corpus contained mostly factual content. However, in the S000 dataset, students were required to demonstrate a critical mind in their writing. Since the S000 dataset was an example of a hard knowledge discipline assignment that required critical writing, and the mark was associated with the rhetorical techniques used rather than just the factual content, it was selected.
2. The essay marks of the S000 dataset were numerical (mark out of 100) unlike BAWE's distinction/merit data. Numerical grades within the dataset allowed better and more precise interpretations, as well as the opportunity to conduct correlational studies in the first instance. This study was conducted to understand whether there is any correlation between the numbers of XIP categories found in each assignment and the marks of these.
3. As S000 launched in 2012 and also ran in 2013, the dataset was sizable, which allowed results from different years to be compared, testing the reliability and generalisability of the method.

5.4.2 Background

S000's assignment required students from the five different scientific strands to work in groups. The aim of the assignment was to work in multidisciplinary teams, developing team-working skills, and more importantly demonstrating skills in written presentation, and posing and evaluating hypotheses.

In this assignment, students were given a scenario (explained below) and expected to write a collaborative scientific proposal (referred to in this thesis as 'team reports') for that

imagined scenario. They were expected to demonstrate a critical voice when writing team reports.

Every team had a 'project leader' who was an Associate Lecturer (referred to in this thesis as a 'tutor'), who also marked the assignment submitted by this group at the end. Tutors had the role of setting out a schedule for working through the tasks, and split the group into two teams, depending on the size if there were 20 or more students in one group. In this dataset, when the teams were examined, it was noted that each team had approximately ten students from all equally distributed five different scientific strands. Therefore, the teams were not significantly different in terms of their size and homogeneity of the various scientific strands within each team.

Within each team, for the team report, each student was required to produce a two-page proposal for the given scenario, which was then ranked by the team members. Teams were required to submit the team report as part of the assignment which contained every student's proposal, written in their ranking order (i.e. the proposal ranked first by the team was listed first in the team report which was then followed by number two, and so on). This team report had a general introduction and conclusion about the team decisions made on the ranking (no more than 22 pages altogether). The following sections explain the assignments, scenarios and assessment nature for the years 2012 and 2013.

5.4.2.1 S000-2012 assignment

The scenario given to students in 2012 required students to work as part of an interdisciplinary team to prepare innovative proposals for practical investigations that might be carried out in support of a human-crewed mission to Mars in 2020.

In 2012, the assignment was divided into three parts:

1. the team's proposed space exploration mission, team reports (33% of the mark)
2. a reflection on the team's performance (25% of the mark)

3. personal reflections (42% of the mark)

For this study, only the first part of the assignment was used because it was the part where student teams were expected to produce argumentative writing. As XIP works well with this type of writing, this part met the requirements. Although the team reports affected 33% of the overall assignment mark, each team report was marked separately out of 100. Therefore, the mark that was used in this study was only the mark of the team report.

5.4.2.2 S000-2013 assignment

In 2013, the given scenario to the students changed. With the main idea remaining the same, this time student teams were expected to write innovative proposals for practical investigations in preparation for the re-colonisation of a remote island on Earth that had been left uninhabited for 13 years.

In 2013, there was a slight change in the nature of the assignment as well. This time the assignment was divided into four parts:

1. critical review of the team report submitted in 2012 (20% of the mark)
2. the team's proposed isolated area exploration mission (25% of the mark)
3. a reflection on the team's performance (25% of the mark)
4. personal reflections (30% of the mark).

In 2013, a new part was added to the assignment and individual students were expected to write critical reviews on the previous year's team. For the correlational study, the second part of the assignment was used (in order to remain consistent with the previous year's dataset), as this was the part where student teams were expected to produce argumentative writing. Each team was marked by the team's own tutor, but all of the tutors used the same assessment rubric that was provided by the module team.

5.4.3 S000 correlational study

As S000 launched in 2012 and also ran in 2013, datasets from both years were used for the correlational study. As the nature of the assignment and assessment rubrics remained the same in both years, there was no expected variability that could affect the results. Although the balance of marks changed, it did not have any effect on the results of this study since only the mark awarded for the team reports was considered, not the overall assignment mark. There were 33 team reports submitted by 275 students in 2012, and 56 team reports submitted by 564 students in 2013, totalling 89 team reports, all of which successfully received at least the required pass mark[9].

5.4.3.1 S000 correlational study results

Correlational studies were conducted separately for the 2012 and 2013 data, and then both together as in both years the idea of the assessment and marking rubrics remained the same. Potential issues related to this are discussed later in the section.

2012 Results

In 2012, there were 33 reports submitted by student teams. The lowest mark was 44 and the highest was 84.8. The mark range was 40.8, the mean mark was 64.2 and the mode and median mark were 64, and the standard deviation was 12.1. As the mode and the mean are the same, it can be inferred that the sample is equally distributed across marks.

A Pearson product-moment correlation coefficient was computed to assess the relationship between the total number of salient sentences found by XIP in a team report and the mark

[9] Note that at the Open University, different modules award different pass grades. Some modules may award a Pass grade only; Distinction and Pass; Distinction, Merit and Pass; or graded credit of pass grades 1, 2, 3 and 4. Therefore merit and distinction terms were not specified throughout the thesis for the OU module dataset. Please see the glossary for the OU terminology.

awarded to the report. The correlation was computed as 0.004 which shows no meaningful correlation.

When correlations between each XIP category and mark were computed, there was a weak positive correlation overall between the team report mark and the total number of TENDENCY sentence types (Pearson's r=0.332). Increases in the number of TENDENCY sentences were correlated with increases in mark. The P-Value is 0.059, which is not significant at $p < 0.05$ level.

Also, there was a weak, negative correlation between the team report mark and the total number of SUMMARY (r=-0.276) sentences found in the report. Increases in the number of SUMMARY sentences were correlated with decreases in mark. The P-Value is 0.12, which is not significant at $p < 0.05$ level.

For the rest of the categories no meaningful correlation was found.

2013 Results

In 2013, there were 56 reports submitted by student teams, which were analysed in the same way as 2012 data. The lowest mark for 2013 was 40 and the highest was 92; which made the range 52.3. The average mark for the assessment was 74.2. The most repeated essay mark, the mode, was 72 and the middle value, the median, was 75.6 and the standard deviation was 11.7 for the 2013 dataset. These numbers were again indicators of a good, equally distributed sample across the marks.

The correlation for 2013 was computed as 0.09; which again shows no meaningful correlation. Overall, there was a weak, positive correlation between the team report mark and the total number of BACKGROUND (r=0.201), SUMMARY (r=0.172), TENDENCY (r=0.146) and NOVELTY (r=0.134) sentences in the report. Increases in the number of BACKGROUND, SUMMARY, TENDENCY and NOVELTY sentences were correlated

with increases in mark. The P-Values for these correlations are 0.137, 0.205, 0.283, and 0.325 repectively, which are not significant at $p < 0.05$ level.

Also, there was a weak, negative correlation between the team report mark and the total number of OPEN QUESTION ($r=-0.237$) sentences found in the report. Increases in the number of OPEN QUESTION sentences were correlated with decreases in mark. The P-Value is 0.08, which is not significant at $p < 0.05$ level.

For the rest of the categories no meaningful correlation was found.

2012 & 2013 combined

When all the marks of the team reports were combined for both years (89 reports in total). The lowest mark was 40 and the highest was 92; which made the range 52.3. The average mark for the assessment was 70.5. The most repeated essay mark, the mode, was 72 and the middle value, the median, was 72 and the standard deviation was 12.9 when the 2012 & 2013 datasets combined.

With both S000 datasets together, the correlation between the essay mark and the total number of salient sentences identified by the XIP computed as 0.128, a weak positive correlation. The scatterplot below summarises the results. Overall, there was a weak, positive correlation between the team report mark and the total number of salient sentences found in the report. Increases in the number of salient sentences were correlated with increases in mark. The P-Value is 0.232. The result is not significant at $p < 0.05$ level.

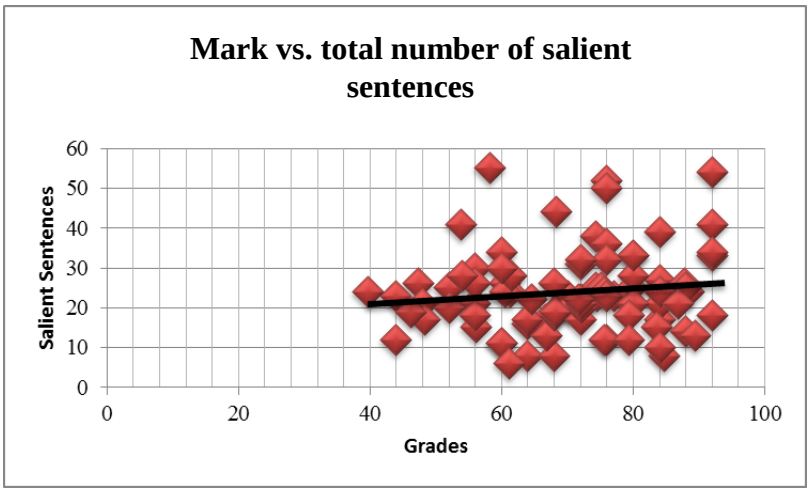


Figure 5.2 Correlation graph showing the relationship between the total number of salient sentences in team reports and mark of the reports in 2012/13

Overall, there was a weak, positive correlation between the team report mark and the total number of sentences that XIP classifies as TENDENCY (r=0.129), SUMMARY (r=0.147), and CONTRAST (r=0.124) in the report. Increases in the number of TENDENCY, SUMMARY, and CONTRAST sentences were correlated with increases in mark. The P-Values for these correlations are 0.200, 0.169, and 0.247 repectively, which are not significant at $p < 0.05$ level.

Also, there was a weak, negative correlation between the team report mark and the total number of OPEN QUESTION (r=-0.167) sentences found in the report. Increases in the number of OPEN QUESTION sentences were correlated with decreases in mark. The P-Value is 0.117768. The result is not significant at $p < 0.05$ level.

For the rest of the categories no meaningful correlation was found.

5.4.4 S000 multiple linear regression study

Correlational studies on the S000 dataset produced some weak correlations. In order to gain a better understanding of how each rhetorical sentence type affects the final mark of the team report, a 'generalised' multiple linear regression study was carried out. As the dependent variable mark had numerical values (unlike BAWE's dichotomous

representation), generalised multiple linear regression was used instead of the binary logistic regression.

5.4.4.1Results

When the results were computed using SPSS, this time the R-square was given as an adjusted R-square format rather than Nagelkerke's R-square. This is because the latter is used for binary logistic regression but the former is used for the interpretation of the generalised multiple linear regression models. Adjusted R-square measures the proportions of the total variability in the dependent variable that is explained by the independent variables of the model. For this model the R-square was -0.005. This negative result indicates that the model has no statistical value and therefore the coefficients should not be interpreted. The null hypothesis was also checked. This is concerned with whether the model has an explanatory power or not; in other words, whether all the coefficients on the independent variable are zero or not. That is the same as saying that none of the independent variables help to predict the value of the dependent variable. To answer whether the null hypothesis should be rejected or not, the p value was checked. The p value for the model was 0.486 which is very weak. It means that there was at least a 51.4% chance that the model supports the hypothesis based on the sample data. The hypothesis was that independent variables, XIP categories, have an explanatory power on the dependent variable mark. However, there could be up to 48.6% chance that the apparent support is due only to random chance which is a very weak standard. In other words, there is no strong evidence to reject the null hypothesis.

Overall, the generalised multiple linear regression analysis showed no meaningful results to interpret, as the adjusted R-square was found to be negative. R-square compares the fit of the chosen model with that of a horizontal straight line (in other words, the null hypothesis). R-square is negative only when the chosen model does not follow the trend of the data, so fits worse than a horizontal line.

5.4.5 Discussion of S000 Analysis

The S000 dataset allowed correlational studies unlike the BAWE corpus which does not give the essay marks as numerical values but with distinction and merits. After conducting correlational studies with team reports from each year, the 2012 and 2013 team reports were collated and the study was repeated. This was possible as the marking rubrics remained the same which eliminates the instrumentation biases. However, although the nature of the assignment was very similar in both years along with the assessment rubrics, different teams were working on different scenarios. This creates a challenge in terms of reliability because changes in mark might be affected by the topic of the scenario and the background information provided to the teams. For instance, the 2013 teams were given an example report from the previous year. Therefore, changes in marks cannot solely be explained by the better quality of writing because of other possible variables such as the topic difficulty and the quality of the material provided to students.

Since the correlational studies yielded weak results, in order to gain a better understanding of how each XIP sentence type affects the essay mark, multiple regression analysis was carried out, but the model had no explanatory power on mark. S000's multiple regression analysis result aligns well with the BAWE's Physical Sciences regression model. Therefore, it can be concluded that XIP does not perform well enough on student essays written for hard disciplines.

Typically, as stated in the literature review, hard knowledge field writing at university level does not actively encourage argumentation. Although S000 was an undergraduate science module, it was a multidisciplinary science module requiring its students to produce critical, argumentative writing. It is therefore expected that argumentation would be picked up by the XIP, which it did indeed do but it was not found to be effective on the essay mark for the S000 sample.

Prior to the statistical analysis, false hits from the XIP output were eliminated. Bibliography sections were discarded from all of the texts as some of the titles were identified as salient by the XIP. Similarly, non-existence of the line breaks between figure/table headings and the following sentences meant the following sentence was picked up as salient although it was not. The reason for this was that those figure titles had linguistic features, meta-discourse, that XIP could identify as salient. Therefore, certain kinds of false hits, that would weaken the statistics, were excluded in order to obtain a more precise test of whether XIP's categories relate with quality. Moreover, additional analysis of the false negatives was conducted, in order to make sure the XIP tool does not miss any rhetoric. There were no false negatives found within the dataset.

Since the XIP output was eliminated from such false hits, one possible reason for not finding any statistical evidence that the XIP categories were not effective on the essay mark could be that the linguistic features used by these students did not align with the linguistic features that XIP looks for. To investigate this further, the accuracy of the XIP results were examined from all the student writing. This showed that the accuracy of these results was not flawed as there was not a single poor identification. All the sentences were accurately picked up as salient under accurate XIP categories. The following examples from the student texts for each XIP category illustrate this:

- **EMPHASIS**

  "It is believed this is a vital study for the success of further long stay missions and that at this important stage of exploration all precautions must be taken to protect both planets biospheres prior to the possible habitation of Mars."

- **BACKGROUND**

  "Previous NASA research has shown there is enough carbon dioxide on Mars to support microbial life (McKay et al, 2001)."

- **CONTRAST**

  "The 1976 twin Viking landers provided only evidence that Mars was an inhospitable, sterile planet (Smith, 2011) however the missions since then have proven Mars to be a more complex and interesting planet than we previously knew."

- **SUMMARY**

  "This report sets out several possibilities for power on a mission to Mars; it states the information that will be required to decide on a source before the mission departs."

- **NOVELTY**

  "This investigation can add further evidence to any results found or if no evidence of life is found then this would be an opportunity to investigate a wider range of possible environments using and improving previous methods."

- **TENDENCY**

  "Not only will important questions about the planet be answered, but it could also serve as a starting point for future manned missions, and would create a huge increase in potential investigations that could be carried out on the Martian planet."

- **SURPRISE**

  "Data suggests that most of Mars' crustal rocks are mafic in composition, specifically basaltic, with evidence of some quartz - containing felsic igneous rocks depending on their molecular structure, varying minerals emit long-wave TIR (thermal infra-red region) radiation differently, so that they have distinct spectral patterns. These rocks were compositionally similar to each other and to terrestrial andesite, but unexpectedly high in silica and potassium, and low in magnesium compared to Martian soils."

Since the false hits were eliminated and the accuracy of the XIP results was controlled, the S000 study triangulates the BAWE study results, and suggests that the XIP did not work well with undergraduate science writing, especially in hard disciplines. The research question can be answered to say that the automated rhetorical parser XIP cannot be used to identify good indicators of academic writing in undergraduate student writing for hard knowledfe fields and so no further repetition is needed with a different dataset.
To answer the extent of how XIP can be used, the analysis however should be repeated on student writing from other disciplines, a complementary social sciences or humanities discipline (e.g. educational psychology, social sciences etc.). The next sections describe the studies repeated with student writing from these disciplines, in which students' ability to write critically and argumentatively was measured and valued.

5.5 E000 DATASET

5.5.1 Introduction

The student writing in this dataset came from one of the third-level undergraduate education and arts modules of The Open University, which aligns with the arts and humanities BAWE category. The reason for selecting this dataset was the nature of the assignment, given in learning outcomes and marking rubrics used by the tutors. Students were expected to read academic texts critically, identify and evaluate positions and arguments from those, synthesise that information, and analyse and critically evaluate others' arguments and evidence in order to reach their own conclusions. The statistical studies were conducted to understand how XIP performs within soft disciplines as this sort of discipline is the one most likely to award marks for rhetoric, unlike hard disciplines.

5.5.2 Dataset

In E000, a course on literature for young people, students study key examples of novels, picture books, poems and creative performance produced for children aged between 3 and 18. Students read a selection of related critical material and consider major themes, issues and debates in the field such as the relationship of children's literature to the conceptions of childhood, the question of whether children's literature should instruct or delight, and the tension between popular and prestigious literature for children. Throughout the module, students write six essays, each of 2000 words, about these themes. At the end of the module, students are expected to produce an individually written, longer essay (3000 words) which is the examinable component of the module; that means it takes the place of an examination. Unlike the previous six essays, the end of module assessment aims to assess the knowledge, understanding and skills students have developed throughout the module as a whole.

5.5.3 E000 assignment

In the assignment, students were required to answer the question: "Does children's literature always have an overt and/or a hidden agenda? Discuss this question with reference to the chapters by Kimberley Reynolds and Peter Hunt in Reader 1, Approaches and Territories, together with at least three Set Books from two or more blocks of E000." The assignment required students to engage in depth with texts and approaches explored within the module. Students were expected to engage in critical thinking and in evaluating the materials they had studied. They were assessed, in part, on their ability to think through the strengths and limitations of the materials they used, and to express this critical thinking clearly in their writing. Students were expected to demonstrate that they could sift material and present points so as to set out a coherent argument, garnering support from the sources

and positions represented across the module. In terms of learning outcomes (see APPENDIX B for full list), students were expected to:

- read academic and other texts critically, identifying and evaluating positions and arguments
- synthesise information and ideas from different sources, in order to reach their own conclusions
- analyse and critically evaluate arguments and evidence, from a variety of sources, relating to particular theoretical and analytical approaches to the study of children's literature.

5.5.4 Assessment

For this assignment 1307 students each submitted an essay. Each of these essays was marked out of 100 and those that received a mark below 40 failed.

Each essay was marked by the tutors. Each tutor marked around 80 essays and all tutors used the same marking guidelines and rubrics (see APPENDIX C and D) provided by the module team. Based on these guidelines, tutors marked the student writing in six points:

1. relevance to question
2. knowledge and understanding
3. approach to alternative explanations and arguments
4. construction of argument
5. clear expression and use of academic conventions
6. skills of independent study.

Unlike S000, the mark distribution for each of these points was not made clear to the tutors. Therefore, what each of these points would mean numerically is not information that was available.

5.5.5 E000 correlational study

A correlational study was conducted to assess the relationship between the total number of salient sentences found by XIP in student essays and the mark of these essays. The correlation was computed as r=0.190, which means there was a weak, positive correlation between the essay mark and the total number of salient sentences extracted by XIP. The P-Value is < 0.001. The result is significant at p < 0.05 level.

Increases in the number of salient sentences are correlated with increases in mark.

Table 5.7 Correlational Study Results for each rhetorical sentence type

Rhetorical sentence type	Value of the correlation coefficient	Strength of the correlation	Significance at P<0.05 level
CONTRAST	r=0.151	Weak	P < 0.001
BACKGROUND	r=0.109	Weak	P < 0.001
TENDENCY	r=0.025	No meaningful correlation	P= 0.366
EMPHASIS	r=0.076	No meaningful correlation	P= 0.006
NOVELTY	r=0.097	No meaningful correlation	P= 0.004
SUMMARY	r=0.005	No meaningful correlation	P= 0.857
QUESTION	r=0.007	No meaningful correlation	P= 0.800

When correlations between each XIP category and mark were computed, no negative correlation was found overall, and no meaningful correlation was found with the sentence types: TENDENCY, EMPHASIS, NOVELTY, SUMMARY and OPEN QUESTION. There was a weak, positive correlation between the essay mark and the total number of CONTRAST (r=0.151, P < 0.001) and BACKGROUND (r=0.109, P < 0.001) sentences: increases in the number of these sentence types were weakly correlated with increases in mark.

5.5.5.1 Correlational study discussion

The correlational study with E000 did not show any promising results except weak positive correlation between the total number of salient sentences in essays and the marks given, as well as weak positive correlations between the rhetorical sentence types CONTRAST, BACKGROUND and marks, and no negative correlation was identified.

Although weak correlations were found between the essay mark and the total number of sentences, statistically it may simply be that higher grade texts were longer, and the more XIP sentences they might have had. However, all the assignments that were marked above pass mark had around 3000 words, which was a fixed word limit that most of the students adhered to. The assignments that failed the assessment had a relatively lower number of sentences which could have caused this result. Since the correlation results were weak even to make such conclusions, these assignments were included in the dataset.

Whereas in the great majority of the essays the grade was correlated with the number of salient sentences detected by XIP, in some rare cases high grades were given by the tutors to essays with very few salient sentences, and conversely, low grades were given to essays with a relatively greater number of salient sentences. A close look at some of these essays provided some insight into these cases.

High graded essays with few salient sentences had a strikingly vivid and literary style, which did not strictly follow the patterns of concise scholarly communication. These essays conveyed a personal approach, showed deep knowledge, and used unconventional expressions. Alternative explanations required by the marking grid were provided, however they were embedded into a particular narrative flow, in which the expression of contrast was distributed throughout several sentences (underlined). Consider the following extract:

> *"As Hunt states 'sameness and difference is the essence of children's books; they have recurrent ideas' (2009a,p. 71). He goes on to cite … [a list of examples.] But is this the only tradition the book breaks? Based upon the themes detailed above this essay will look at what similarities and differences A*

Monster Calls has to children's literature from the last 250 years, focusing particularly on Tom's Midnight Garden."

Instead of referring to the alternative arguments through expressions such as 'contrasting analyses' or 'critical debates', the author of this essay lays them out in several steps. What was also observed in the case of low-graded essays containing a relatively high number of salient sentences is that in comparison, their style was simple and schematic, and sometimes their syntactic structure was not clear:

"I do not think any of the themes I have mentioned were written about to change or challenge aspects of the community, I believe these issues were just to define the culture of society as it was in the Victorian era and to reinforce the roles subliminally."

Regarding the correlations between the XIP categories and the essay mark, XIP would not be expected to have something to say about all possible measures of quality, which in this case was marked in six points. Tutors marked these assignments based on these six criteria, and not all of these are relevant to XIP. However, what each grading criterion's value was numerically, i.e. what percentage each of these points represented of the whole essay mark, is unknown. Therefore, although it would be more plausible to establish the correlation between each XIP category and each grading criterion that potentially match (possibly the criteria 3, 4, and 5 as little correlation would be likely with 1,2 and 6), it was not possible to do so.

Although these results gave some insights about the correlation between XIP findings and marks, they do not tell the whole story. The next section describes the multiple regression analysis by using the 1307 essays in the E000 dataset to interpret how strongly each sentence type affected the final mark.

5.5.6 E000 multiple linear regression

At the start, the regression model included the *total number of salient sentences* as an independent variable, in addition to the *total number of sentences for each rhetorical*

sentence type. However, the tests to see if the data met the assumption of collinearity indicated that multicollinearity was a concern. (Multi)collinearity refers to the case when there are very high correlations between independent variables in the regression model (Haitovsky, 1969). Due to multicollinearity, independent variables might seem not significant when they actually are. Unlike the previous regression analyses with the S000 dataset, this model did not satisfy the collinearity assumption. Therefore, 'the total number of salient sentences' independent variable was removed from the model as it was strongly correlated with CONTRAST (p=0.635) and moderately correlated with SUMMARY (p= 0.481) and EMPHASIS (p= 0.401). Therefore, this study regressed the mark on each XIP category.

In the multiple linear regression model, the mark of the essays was taken as the dependent variable and the number of salient sentences for each XIP category (TENDENCY, EMPHASIS, NOVELTY, SUMMARY, OPEN QUESTION, CONTRAST and BACKGROUND) marked up in the essays as independent variables.

The regression model proved to be highly significant. Following normal convention, $p \leq 0.05$ signifies a statistically significant result, and $p \leq 0.001$ is regarded as highly significant. The p value for this model was less than 0.001, which indicates that the model is statistically highly significant. It means that this is strong evidence to investigate further how strongly independent variables help to explain the essay mark with the model.

For this model the adjusted R-square was 0.048, which means that 4.8% of the total variability in mark was explained by the independent variables.

When each independent variable was analysed, it was found that the two of the independent variables: CONTRAST and BACKGROUND were statistically highly significant and have explanatory power for the dependent variable essay mark (CONTRAST, $p \leq 0.001$; BACKGROUND, $p \leq 0.001$).

When unstandardised coefficients were examined for these two independent variables, the following interpretations were made:

- *for a one unit increase in the number of CONTRAST sentences within essays, the model predicts that the dependent variable, the essay mark, will increase between 0.498 and 1.078 points (calculated as B±2*Std. Error), holding all other independent variables fixed/constant.*
- *for a one unit increase in the number of BACKGROUND sentences within essays, the model predicts that the dependent variable, the essay mark, will increase between 1.075 and 3.431 points, holding all other independent variables fixed/constant.*

For the rest of the independent variables, the p value was not significant and therefore they could not be interpreted in the same way as CONTRAST and BACKGROUND.

5.5.7 Discussion of E000 analysis

The tutors' marking grid contained criteria for evaluating the essays according to six aspects, two of which are particularly in line with XIP's framework: "Approach to alternative explanations" and "Construction of academic argument". Thus, it is most likely that these two aspects underlie the correlations between the tutors' marks and XIP results on sentences labelled as CONTRAST and BACKGROUND. Sentences labelled CONTRAST capture the expression of tensions, contrasts between ideas, models or research directions, whereas the sentences labelled BACKGROUND make reference to relevant other work which is considered consensus knowledge. Thus these two sentence types in XIP do indeed perform discourse functions that convey "alternative explanations", which in turn are organic parts of "academic argument".

The quantitative study did not show any statistically significant correlations between the marks and the other salient sentence types detected by XIP: SUMMARY, EMPHASIS, NOVELTY, TENDENCY and OPEN QUESTIONS. Taking into account the evaluation aspects and the object of the essays, the following explanation can be proposed.

The SUMMARY sentences merely convey the idea that the author summarises her essay. Thus these sentences do not contribute to any of the evaluation aspects. Referring to new research directions (NOVELTY), raising open questions, emphasising ideas as surprising, or important, and describing research tendencies are not usual discourse moves in literature analysis at undergraduate level; these are elements of argumentation schemes in mainly empirical research.

Whilst salient sentences do indicate the author's awareness of alternative analyses, and show efforts to develop scholarly argumentation, their mere presence does not imply that the alternative analyses are discussed at a sufficient level, or that the argumentation is sound, well-structured or coherent. Their presence simply signals that the writer does include some content on alternative analyses, and that this argumentation does treat the topic in a scholarly style. Still the fact that the number of salient sentences shows a correlation with the marks indicates that the more scholarly meta-discourse is present in a student essay the more likely it is that it gets a better mark in the evaluation.

In order to interpret the performance of the XIP, 225 automatically detected salient sentences were evaluated, and 49 (22%) of them did not play the role of the scholarly argumentation in the essay. An important source of errors is related to the specificity of literary essays that the current version of XIP does not account for. Since these essays involve the analysis of literary work, the rhetorically salient expressions may also be part of that analysis and not of the scholarly argumentation. The following sentence, which refers to the children's story, Peter Pan, illustrates such a non-rhetorical expression detected by XIP (underlined):

> *Wendy <u>is not seen to challenge this role</u> even when she is out of her comfort zone and enters Neverland.*

Since the XIP has not been adapted to work in this particular domain, such issues are expected. However, the noise in the literary essays in this study does not amount to a proportion that would undermine the validity of the statistical correlations.

An internal validation was carried out using a randomly selected subset of the overall data. IBM's SPSS statistical software package was set to select half of the data randomly, and ran the regression analysis on this. This produced exactly the same results: the CONTRAST and BACKGROUND categories proved to be highly significant, and have an explanatory power on the dependent variable, the essay mark. For external validation, the next chapter describes the study with a similar dataset, L000.

5.6 L000 DATASET

5.6.1 Introduction

The previous study conducted with E000 data produced promising results and allowed valid interpretations regarding XIP's performance. For external validation, a similar dataset, L000, was used to explore whether similar results would be found with a different dataset. The first reason for selecting this dataset was because it was similar to the E000 dataset. Second, the assignment that students worked on was specifically about argumentation. Students were taught about argumentation, and then produced an argumentative piece of critical writing. Finally, the marking guidelines suggested that certain correlations would appear with specific XIP categories.

5.6.2 Dataset

The student writing in this dataset came from one of the first level undergraduate education and arts modules of The Open University, which aligns with the Arts and Humanities BAWE category. L000, a course on English for academic purposes, aims to develop the communication skills of its students who need them for successful academic study in English. Throughout the module, students work with a wide range of texts from different subject areas to develop the academic English that is most relevant to their study. The intention is that students explore ways of reading and writing academic texts. Students

develop academic listening and speaking skills in an online environment and explore ways of reading and writing academic texts.

Specifically, in the module, for five weeks, students study how to construct arguments in written academic texts. Students learn about organising academic arguments in assignments, designing arguments that have a single point of view, and then explore designing arguments that have more than one point of view. After five weeks, students write individual critical discussion essays, which are 800 words long for revising what they have learned about argumentation. At the end of the module, students are expected to produce an individually written longer assignment (1000 words) which is the examinable component of the module; that means it takes the place of an examination. Both assignments were chosen as datasets for the study.

5.6.3 L000 assignments

The 800-word assignment required students to write a critical discussion essay based on the materials already studied in the module. Students had to undertake the following task: *"Rainforests have been declining rapidly over the last few decades. There are various factors responsible for this decline, resulting in serious impacts in the environment and economy. Critically discuss the causes of deforestation and solutions to it."* While doing this, students were expected to engage in depth with three texts explored within the module. Students were expected to engage in critical thinking and in evaluating the materials they had studied. They were assessed, in part, on their ability to use source materials, include background information about the issue, outline causes and possible solutions by summarising, paraphrasing, and referencing.

In the end of module assignment, the same students wrote a similar but longer critical discussion essay in response to the more demanding task*: "Drawing on any five of the ten source texts, discuss the relative importance of 'nature versus nurture' as predictors of happiness"*.

5.6.4 Assessment

For each assignment, 343 students each submitted an essay. Each of these essays was marked out of 100 and those that received a mark below 40 failed. Each essay was marked by tutors, who used the same rubric provided by the module team. This rubric was produced centrally and was used by everyone marking on the presentation of L000 that began in October 2012.

Along with the rubric, the tutors were given a marking grid in which they assessed five aspects of their students' writing:

- use of source material (25 points)
- design and development of the text (25 points)
- academic writing style (25 points)
- grammatical correctness (20 points)
- qualities of presentation (5 points)

5.6.5 L000 multiple regression analysis

Two regression models were produced for each assignment. In both multiple regression models, the marks of the essays were taken as the dependent variable regressed by the total number of salient sentences for each XIP category, independent variables.

For the **shorter assignment**, the regression model proved to be highly significant ($p \leq 0.001$). This means that there was very strong evidence to reject the null hypothesis so that the model gave an opportunity to investigate further how strongly independent variables help to explain the essay mark. For this model, the adjusted R-square was 0.06, which means, when converted into percentages, 6% of the total variability in mark was explained by the independent variables. When each independent variable was analysed, it was found that the null hypothesis could be rejected for two of the independent variables:

SUMMARY and BACKGROUND. These two categories were highly significant and had explanatory power for the dependent variable essay mark (SUMMARY, $p \leq 0.001$; BACKGROUND, $p = 0.031$).

When unstandardised coefficients were examined for these two independent variables, the following interpretations were made:

- *for a one unit increase in the number of SUMMARY sentences within essays, the model predicts that the dependent variable, the essay mark, will increase between 1.237 and 4.077 points (calculated as B±2*Std. Error), holding all other independent variables fixed/constant.*
- *for a one unit increase in the number of BACKGROUND sentences within essays, the model predicts that the dependent variable, the essay mark, will increase between 0.637 and 16.137 points, holding all other independent variables fixed/constant.*

For the rest of the XIP categories, the independent variables TENDENCY, EMPHASIS, NOVELTY, CONTRAST and OPEN QUESTION, the p value was not significant, therefore they could not be interpreted in the same way as SUMMARY and BACKGROUND for the shorter assignment.

When the **end of module assignment essays** were used as a dataset, the regression model proved to be highly significant again ($p \leq 0.001$). For this model the adjusted R-square was 1.30, which means, when converted into percentages, 13% of the total variability in mark was explained by the independent variables. When each independent variable was analysed, it was found that the null hypothesis could be rejected for three of the independent variables: SUMMARY, CONTRAST and EMPHASIS. These three categories are highly significant and have explanatory power for the dependent variable essay mark (SUMMARY, $p \leq 0.001$; CONTRAST $p \leq 0.001$; EMPHASIS $p = 0.003$).

When unstandardised coefficients were examined for these two independent variables, the following interpretations were made:

- *for a one unit increase in the number of SUMMARY sentences within essays, the model predicts that the dependent variable, the essay mark, will increase between 1.681 and 3.749 points (calculated as B±2*Std. Error), holding all other independent variables fixed/constant.*
- *for a one unit increase in the number of CONTRAST sentences within essays, the model predicts that the dependent variable, the essay mark, will increase between 0.002 and 2.309 points (calculated as B±2*Std. Error), holding all other independent variables fixed/constant.*
- *for a one unit increase in the number of EMPHASIS sentences within essays, the model predicts that the dependent variable, the essay mark, will increase between 0.556 and 2.848 points, holding all other independent variables fixed/constant.*

For the rest of the XIP categories, the independent variables TENDENCY, NOVELTY, BACKGROUND and OPEN QUESTION, the p value was not significant; therefore they cannot be interpreted in the same way as SUMMARY, CONTRAST and EMPHASIS for the end of module assignment.

5.6.6 Discussion of L000 analysis

The tutors' marking guidance notes contained criteria for evaluating five aspects of the essays, and of these; "Design and development of the text" was particularly well aligned with the XIP framework. The guidance notes under this aspect tell tutors that they should expect their students to:

- give general background to the topic (BACKGROUND)
- identify key concepts in the essay (EMPHASIS)
- say how the essay will be organised (SUMMARY)

- use connectives to introduce challenges such as ‘however’, ‘on the other hand’ (CONTRAST).

Next to each bullet point above, matching XIP categories were added in parentheses; they were not specified in the marking guidelines. As can be clearly seen, these expectations underlie the correlations between the tutors’ marks and XIP results on sentences labelled as BACKGROUND, EMPHASIS, CONTRAST and SUMMARY. Sentences labelled as SUMMARY summarise the goals or results of the essay, EMPHASIS emphasises the importance of ideas, CONTRAST captures the expression of tensions, contrasts between ideas, models or research directions, whereas the sentences labelled as BACKGROUND make reference to relevant other work. Thus these four sentence types in XIP do indeed perform discourse functions that tutors should expect from their students.

Additionally, there was a clear improvement in students’ writing which is caught by the XIP. Since in the first assignment students only discussed the reasons for deforestation, BACKGROUND and SUMMARY sentences came up in the regression model. In the end of module assignment, however, students were expected to compare opposing views on what makes people happy: nurture or nature. Therefore, the sentence type CONTRAST came up in the regression model along with EMPHASIS and SUMMARY. This result shows an obvious improvement in students’ argumentative writing skills that is also caught by XIP.

5.7 Chapter conclusion

This chapter addressed the first subsidiary research question of this thesis: "To what extent can the automated rhetorical parser XIP be used to identify indicators of good academic writing in undergraduate student essays from different disciplines, as judged by the essay grade?"

XIP was designed to work on peer-reviewed academic research writing; but by a team with no training in education, or intent for it to be used in education. However, it connects with education to the degree that there is overlap in the hallmarks of research articles and the kinds of writing that academics seek to nurture in undergraduate students, and reward through grading criteria, which is what this study has dealt with. Therefore, there was a need to understand whether the XIP can be used to identify indicators of good undergraduate student writing. The quantitative data analysis chapter described evaluation studies carried out to test the XIP's performance on undergraduate student essays from various disciplines and levels, using the mark awarded as a measure of the quality of the writing. The studies presented in this chapter sought to assess the quality of the XIP through correlational studies and regression analysis.

The research question can be answered as follows: To a significant extent, depending on the discipline, level and tutors' expectations, the automated rhetorical parser XIP can be used to identify indicators of good academic writing in undergraduate student essays, where these indicators are tested by the essay mark. The following conclusion points can be given based on the studies with different datasets (S000, E000, and L000) and on the BAWE corpus:

- From a learning analytics point of view, it has been found that some of the XIP categories were good predictors of final marks. However, these categories were discipline and level specific.

- Not all of XIP's existing categories were found to have a significant impact on the essay mark. The categories TENDENCY, SURPRISE, NOVELTY and OPEN QUESTION, that are found in journal writing of experienced researchers, did not appear necessary for undergraduate students to get better grades.
- The categories BACKGROUND, EMPHASIS, CONTRAST and SUMMARY, on the other hand, were associated with higher marks.
- XIP was less likely to work well with student writing from hard knowledge fields, whereas XIP performed well with student writing from soft disciplines such as Arts and Humanities.
- XIP did not work for Level 1 student writing, but it was more likely to work at higher levels, Level 2 and Level 3.
- Where tutors' marking guidelines were available to inform the selection of datasets, this served as a better validation of XIP, since it was known that students were being required to produce argumentative writing. When the marking rubric aligned with XIP categories, it was more likely that the presence of some categories correlated with grade. Therefore, it can be argued that XIP was able to detect features of a good advanced student essay automatically in the discipline of the Arts and Humanities.

These promising outcomes suggest that XIP could be used for training undergraduate students and making them aware of these types of categories in order to improve their writing skills as well as to get better grades. However, some of the outliers occurred during the studies have to be acknowledged. Specifically, in the E000 dataset, whereas in the great majority of the essays the grade was correlated with the number of salient sentences detected by XIP, in some rare cases high grades were given by the tutors to essays with

very few salient sentences, and conversely, low grades were given to essays with a relatively greater number of salient sentences.

High graded essays with few salient sentences have a strikingly vivid and literary style, which does not strictly follow the patterns of concise scholarly communication that is used in XIP's algorithm. These essays convey a personal approach, show deep knowledge, and use unconventional expressions, which is why the salient sentences were not picked up. Alternative explanations required by the marking grid are provided; however, they are embedded into a particular narrative flow, in which the expression of contrast is distributed throughout several sentences. Instead of referring to the alternative arguments through expressions such as 'contrasting analyses' or 'critical debates', the author of this example essay lays them out in several sentences.

In the case of low-graded essays containing a relatively high number of salient sentences is that in contrast, their style is simple and schematic, and sometimes their syntactic structure is not clear. The fact, however, that the number of salient sentences shows a correlation with the marks indicates that the more scholarly meta-discourse is present in a student essay, the more likely it is that it gets a better mark in the evaluation. However, these outliers signal the fact that XIP requires some alterations which need to be explored within the following studies given in next chapters. Based on this chapter for instance, sentence-based analysis could be spread to paragraph-level so that when an author lays expressions across several sentences, this could be captured.

The quantitative data analysis studies advance the understanding of the relevance of XIP's rhetorical parsing for undergraduate writing. There are better answers to the research question: *"To what extent can the automated rhetorical parser XIP be used to identify indicators of good academic writing in undergraduate student essays from different disciplines, as judged by the essay grade?"* On the other hand, it cannot be said that these answers are completely true, as there is an extent to which other aspects also come into

play. Therefore, while for many educators the statistical correlation with grade is an important question to answer, before such a parser can be considered as a practical tool, it requires validation by tutors themselves. The next chapter describes the qualitative data gathered by consulting tutors to gain a better understanding of their views on what makes good student writing.

6

ONE-TO-ONE INTERVIEWS WITH MARKERS

6.1 Introduction

This chapter addresses the second subsidiary research question: "How do educators define the attributes of good student writing, and to what degree can the automated rhetorical parser, XIP, identify the presence of these attributes?" Answering this question required an investigation into how educators define the quality of student writing, what they give credit for when marking a student essay, and to what extent the XIP analysis can capture these elements.

The XIP analysis of student writing, explained in Chapter 5, suggested that promising results could be obtained from relating categories used in XIP analysis, to the essay marks for student texts from various disciplines requiring argumentative critical writing, with the exception of hard disciplines, despite the fact that the XIP tool had not been developed for this particular purpose and context. Since it is important to know that this XIP analysis is in line with what educators expect to see in good student writing, it is essential to understand in depth what educators value in writing, and how similar the XIP analysis is with respect to their judgement of quality. The next section of this chapter reports the design details of the study, and how the data were collected, which is then followed by the explanations of the participants. An account of how the data were transcribed and analysed is then given, and, finally, the findings are reported.

6.2 Study

A study investigating how educators judge the quality of students' writing and what makes them give credit for specific features of the text required observation of educators whilst marking an essay. In this study, data was collected by means of individual interviews. Working with educators individually was necessary to ensure that they were in an environment that was close to their natural marking settings which require markers to work alone. Additionally, this would allow them to talk through an essay, using the think-aloud protocol approach (Gambier & Van Doorslaer, 2010, p. 371) to verbalise whatever crosses their mind during task performance, with the sole focus on the student writing and their reasoning, without group interruption or reaction. Therefore, in this study each tutor was interviewed individually. Although the marking guidelines and rubrics that the tutors use could be reviewed, these do not provide a sole basis for the educators' judgement of quality. Therefore, this study required:

- meeting with the OU's E000 educators, who marked an essay during the session (due to the promising results of the multiple regression study, explained in chapter 5, that demonstrated the relation of students' essay marks from E000 with the XIP analysis)
- discussing why they had given credit to certain features
- looking for their definitions of a good essay for a 3rd level arts and humanities discipline.

Moreover, to answer the following part of the research question, - "*to what degree can the automated rhetorical parser, XIP, identify the presence of these attributes?"* -, it is key to determine what they expect from good quality writing and whether their marking overlaps with the XIP analysis. Doing this, required tutors to identify the key sentences that contribute to the final mark, assigned during the session, positively so that the overlapping could be observed, since the XIP works on sentence level. Each tutor was given a student

essay that had received a high grade. The reason for giving each tutor the same essay was to define the common features of good-quality student writing. The student essay had been marked in 2013 using the marking guidelines that were used in this part of the research and had received a high grade when they were submitted. This would allow the understanding of what makes this essay so deserving of its high mark.

The interviews were semi-structured (see APPENDIX F for interview questions). Semi-structured interviews with several key questions help define the areas to be explored, but also allow the interviewer or interviewee to diverge in order to pursue an idea or response in more detail (Gill, Stewart, Treasure, & Chadwick, 2008). Each tutor was talking through an essay, and it required a semi-structure that allowed the flexibility of providing detailed reasons behind their marking and judgement.

For the qualitative research, there is no rule of thumb for how many people need to be interviewed, as this number depends on various factors such as the research study, geographical and cultural limitation on the sample, and the size of the overall number of potential interviewees (Baker, Edwards, & Doidge, 2012). Qualitative research methods differ from quantitative approaches in many aspects, but the latter's emphasis is mainly on numbers capturing information from a wide swathe of dataset (Baker *et al.*, 2012). For qualitative research methods, it is essential to ensure that the participants are the holders of knowledge in the area under investigation. To achieve this, all associate lecturers (ALs) and the module team members who carried out the marking for a course on literature for young people in 2015 were invited to participate in the study. One of the defining factors that would be likely to influence the sample size is the heterogeneity of the population from which the sample is drawn. For this study, the population was all the invited markers, who volunteered to take part in the study (seven volunteers in total). The following table summarises each interviewed tutor, their job title, role, level of expertise, and background.

Table 6.1 Description of interviewed tutors

Name	Job title	Job role	Background
Tutor 1 (T1)	Module Chair	• Lecturer in English Language and Applied Linguistics • Worked as an AL for more than a decade. • Experienced module chair	PhD in academic literacies
Tutor 2 (T2)	Module Chair	• Worked as an AL for more than a decade • Experienced module chair	DPhil in Women Reading Online
Tutor 3 (T3)	AL	• Has worked as an E000 AL since 2009	Undergraduate degree in public libraries, and a Masters' degree in Children's Literature
Tutor 4 (T4)	AL	• Has worked as an AL since 2002 and as E000 AL since 2009	Masters' degree in linguistics, currently doing an EdD in an academic literacies related area
Tutor 5 (T5)	Member of module presentation team	• Has third-marked the scripts • Has monitored the ALs' marking of the assignment	Social Sciences
Tutor 6 (T6)	AL	• Has worked as an AL since 2010 on	English Literature

		various OU modules	
Tutor 7 (T7)	AL	• Has worked as an E000 AL since 2009	PhD degree in Classics

The reason for selecting the E000 module was because the results of the multiple regression study, explained in chapter 5, demonstrated the relation of students' essay marks with the XIP analysis. That study also produced promising results about the XIP's performance on E000 students' essays that were in line with tutors' marking rubrics. Tutors were recruited through the online forum channel of the E000 course page (APPENDIX E). In the end, seven interviews were conducted; each took around 90 minutes. Three tutors, who live close to the Open University campus, were interviewed face-to-face and the rest via Skype. To minimise differences between these methods of data collection, all the sessions were audio-recorded, and the marked essays were collected straight after the session, or via e-mail in the case of the Skype interviews. Observational data, such as gestures or gazes, were not collected; therefore, the different settings did not have any implications for data collection and analysis.

Regardless of whether the interview was carried out face-to-face or via Skype, each session had the same structure:

1. Section one was a general, introductory part of the interview, in which E000 tutors briefly set out their views on assessment and what they felt to be good student writing. This section investigated how these educators defined the quality of writing and its most valuable characteristics in this context.
2. Section two was the essay-marking exercise. In this section, the tutors were given a student essay and their usual marking rubric (APPENDIX C) to mark the essay. They were specifically asked to highlight the sentences that they thought had a

positive effect on awarding a good mark, not just in terms of writing style but anything that they considered should influence the quality and the essay mark.

3. Section three was a follow-up question and answer session on the highlighting exercise, to discuss specific assessment decisions by the participants and to learn why they had highlighted particular sentences.

The next section begins with the quantitative analysis of the essay marking exercise, section two of the interview, in order to understand what educators value most in student writing, why they give credit to certain features and whether these features can be identified by the XIP.

6.3 Quantitative analysis of the essay-marking exercise

The essay-marking exercise required tutors to highlight the key sentences that they thought would have a positive effect on the final essay mark. They were each given the same E000 student essay from 2013. They were asked to perform regular marking activity with reference to the usual E000 rubric and to talk through their decisions. The essay was five pages long, excluding the bibliography, and contained around 3000 words. There were 88 sentences in total. The XIP identified 33 of these sentences as salient. Salient, key sentences are the landmark sentences like ‘in contrast to REFERENCE, this approach uses...’.

Taking the XIP’s and the tutors’ highlights, and the similarity and overlap between tutor pairings, the XIP and each tutor were then measured using Jaccard similarity index. Similarity could have been calculated using the Pearson coefficient (cosine similarity). However, Pearson assumes the data is parametric, therefore, the distributional assumptions were not met for this dataset. Since the usual p value calculations could not be done and the dataset does not meet the threshold, the Pearson coefficient was not used. There are many similarity indices, and all have claimed advantages and disadvantages. Jaccard’s similarity coefficient is used to compare similarity statistically between finite sample sets.

Since there is no advantage of using one over another in this context, Jaccard's similarity index was chosen. The similarity and overlap are important in terms of the second subsidiary research question for two reasons. First, when these measures are investigated between tutors, they will help to understand the reliability of the marking. Second, when these measures are compared between tutors and the XIP, it will help to understand to what degree XIP can identify the presence of these attributes. The Jaccard similarity index is calculated as follows:

- Suppose the comparison is between two tutors: T1 and T2.
- The number of highlighted sentences is counted for each tutor, say those numbers are A and B.
- Then the number of sentences highlighted by both is counted, say this is C.
- Then the Jaccard Index (J) is: C/(A+B-C)

In order to measure the significance of this result, J, Real (1999) tables of significant values for the Jaccard similarity index were used. The first table in Real (1999, p. 31) provides 'critical values of J' with a probability level P. In this context, 'critical' means the numerical values that should be achieved to satisfy significance. To get a significantly high correlation with $P>0.05$, Real's Table 1 provides the associated N value. The N value equals (A+B-C). Other tables produced by Real provide the Jaccard similarity index, N value and the significance test results for each case. To illustrate this point, assuming Tutor 1 (T1) highlighted 9 sentences, the XIP picked up 9 , of which 6 had been highlighted by Tutor 1:

- The N value is = 9+9-6 = 12
- The J value is = 6/12 = 0.50
- N is 12, and J is 0.50, but to get a significantly high correlation with $p<0.05$, the J index needs to be 0.667 or greater (that is, the value in the row N=12 and the

column J+(0.05) in Table 1 of Real's paper). In this example, the similarity between T1 and the XIP therefore is not significant.

The following table shows the total number of sentences highlighted by the seven tutors and the number of common highlights between the XIP and each tutor.

Table 6.2 Total numbers of sentences highlighted by each tutor and common sentences with the XIP

	Total number of sentences highlighted by tutor	Total number of common sentences highlighted by the XIP and the tutor
T1	37	25
T2	13	7
T3	45	22
T4	32	15
T5	25	8
T6	16	5
T7	28	14

The Jaccard similarity index results for all possible pairings are given in the following table with calculations. The significance test results between the XIP and each participant are also shown in this table. The required J values are obtained from Real's (1999, p.31) Table 1 based on the associated N value.

Table 6.3 Jaccard similarity index for each tutor and the XIP

Tutor & XIP pairings	Jaccard similarity index	N	Required J	Significance test
T1 & XIP	25/(37+33-25)=**0.55**	45	>=0.4667	**Highly significant** (p<=0.01)
T2 & XIP	7/(13+33-7)=**0.18**	39	>=0.4872	Not significant

T3 & XIP	22/(45+33-22)=**0.39**	56	>=0.4643	Not significant
T4 & XIP	15/(32+33-15)=**0.30**	50	>=0.46	Not significant
T5 & XIP	8/(25+33-8)=**0.16**	50	>=0.46	Not significant
T6 & XIP	5/(16+33-5)=**0.11**	44	>=0.4773	Not significant
T7 & XIP	14/(28+33-14)=**0.29**	47	>=0.4681	Not significant

The Jaccard analysis results showed a highly significant similarity between the highlights of the XIP and the first tutor; and no significant similarity between the marks of XIP and those of the other six tutors.

The Jaccard analysis was also performed between tutors to find out whether they agreed with each other and if their marking was similar. The following matrix shows the common number of sentences highlighted by tutors.

Table 6.4 The common number of sentences highlighted by tutors

	T1	**T2**	**T3**	**T4**	**T5**	**T6**	**T7**
T1	-	8	24	20	10	7	18
T2	8	-	8	8	4	2	5

T3	24	8	-	19	16	8	15
T4	20	8	19	-	10	9	12
T5	10	4	16	10	-	6	8
T6	7	2	8	9	6	-	2
T7	18	5	15	12	8	2	-

The Jaccard similarity coefficient is calculated as before and the following table demonstrates the results, highlighted with yellow, for each tutor pairing.

Table 6.5 Jaccard similarity index coefficient matrix for each tutor pairing

Jaccard similarity coefficient	**T1**	**T2**	**T3**	**T4**	**T5**	**T6**	**T7**
T1							
T2	8/(37+13-8) = 0.1904						
T3	24/(37+45-24) = 0.4137	8/(13+45-8) = 0.16					
T4	20/(37+32-20) = 0.4081	8/(13+32-8) = 0.2162	19/(45+32-19) = 0.3275				

T5	10/(37+25-10) = 0.1923	4/(13+25-4) = 0.1176	16/(45+25-16) = 0.2962	10/(32+25-10)= 0.2127			
T6	7/(37+16-7)= 0.1521	2/(13+16-2)= 0.074	8/(45+16-8) = 0.1509	9/(32+16-9)= 0.2307	6/(25+16-6)= 0.1714		
T7	18/(37+28-18)= 0.3829	5/(13+28-5)= 0.1388	15/(45+28-15) = 0.2586	12/(32+28-12)= 0.25	8/(25+28-8)= 0.1777	2/(16+28-2) = 0.047	

The next table shows the pairwise significance test results. The N value, the required J (RJ) value for significance and the result is given as follows.

Table 6.6 Pairwise significance test results between all tutors

N, RJ & J Sig.	**T1**	**T2**	**T3**	**T4**	**T5**	**T6**	**T7**
T1							
T2	N=37+13-8= 42 RJ>=0.4762 **J=0.1904 Not significant**						
T3	N=37+45-24= 58 RJ>=0.4483 **J=0.4137 Not significant**	N=13+45-8= 50 RJ>=0.4600 **J=0.16 Not significant**					
T4	N = 37+32-20= 49 RJ>=0.4694 **J=0.4081 Not significant**	N=13+32-8= 37 RJ>=0.4865 **J=0.2162 Not significant**	N=45+32-19= 58 RJ>=0.4483 **J=0.3275 Not significant**				
T5	N=37+25-10= 52 RJ>=0.4615 **J=0.1923 Not significant**	N=13+25-4= 34 RJ>=0.5000 **J=0.1176 Not significant**	N=45+25-16= 54 RJ>=0.4630 **J=0.2962 Not significant**	N=32+25-10= 47 RJ>=0.4681 **J= 0.2127 Not significant**			

T6	N=37+16-7=46 RJ>=0.4783 **J=0.1521** **Not significant**	N=13+16-2=27 RJ>=0.5185 **J=0.074** **Not significant**	N=45+16-8=53 RJ>=0.4528 **J=0.1509** **Not significant**	N=32+16-9=39 RJ>=0.4872 **J=0.2307** **Not significant**	N=25+16-6=35 RJ>=0.4857 **J=0.1714** **Not significant**		
T7	N=37+28-18=47 RJ>=0.4681 **J=0.3829** **Not significant**	N=13+28-5=36 RJ>=0.5000 **J=0.1388** **Not significant**	N=45+28-15=58 RJ>=0.4483 **J=0.2586** **Not significant**	N=32+28-12=48 RJ>=0.4583 **J=0.25** **Not significant**	N=25+28-8=45 RJ>=0.4667 **J=0.1777** **Not significant**	N=16+28-2=42 RJ>=0.4763 **J=0.047** **Not significant**	

According to Jaccard analysis results, there are no significant similarities between any of the tutor pairings.

6.4 Discussion of the Jaccard analysis results

According to the Jaccard analysis results, the highlighting carried out by each tutor was significantly different to that of the others. The assumption had been that tutors would share the same understanding about what makes good-quality student writing, so their highlights would be similar, and the overlap between the XIP and the tutors could be measured reliably. However, this proved not to be the case.

There could be various explanations for this result. Considering that all these participants had more than five years' experience of marking such an EMA using the same marking scheme, one explanation could be that human marking is not reliable. This essay was marked earlier in 2013 by two ALs and a third marker. The essay grade was agreed as 92, pass one, in coordination meetings (see GLOSSARY) and the Open University approved granting this mark to the essay. In this case, the expectation was that all the tutors would award an essay mark in the high 80s, or low 90s. However, when the tutors were asked to guess the awarded mark, two tutors (T2 and T7) gave marks that were very different to other tutors. Five tutors agreed with the given mark but tutors 2 and 7, both awarded 75, pass two, and reacted negatively to the actual mark (see the following table). Human

marking is not always reliable, which supports the assertion that using automated technologies to support educators' essay assessment processes could be a good idea.

A second explanation could be that the nature of the highlighting exercise was not sufficiently close to their original method of marking an essay. Although tutors were using their usual marking scheme, and were simply asked to highlight the aspects that could make them give positive credit, the results might not clearly demonstrate this. The procedures that were in place during the exercise, such as the unfamiliar process of sentence-by-sentence highlighting and marking, were different and might account for variance in marking. Therefore, it should not be assumed that experienced markers on this course are unreliable, as the university works hard to assure the reliability of the marks assigned. To examine this further, consider the following table. Tutor 2 and 7 estimated the essay mark as 76. It might be expected that they would highlight a similar number of sentences; however, Tutor 2 highlighted 13 and Tutor 7 highlighted 28. This could mean that Tutor 7 undervalued the final mark considering the number of highlights that she thought had a positive impact on the final mark. Alternatively, it could mean that the highlights do not clearly show what she actually valued. Looking at tutor 6 who, with 16 highlights awarded a mark of 87, supports this assertion. The value of R, the Pearson correlation coefficient, between the total number of highlights X = 37, 13, 45, 32, 25, 16, 28 and the estimated essay mark Y = 87, 75, 86, 90, 86, 87, 75 is **0.4308**. This is a moderate positive correlation (p=0.345), which means there is a tendency for the higher number of highlights to be associated with the higher estimate for essay mark (and vice versa). Although technically a positive correlation, the relationship between the variables is moderate to claim this assumption; therefore, conflating the highlighting of sentences and the assigning of a mark would not be helpful.

Table 6.7 Number of highlights, essay mark estimation and reaction to the actual mark for all tutors

Tutor	Total number of highlights	Estimated essay mark	Reaction to the actual mark
Tutor 1	37	85-90	DB: "Okay, so the essay was given 92." T1: "Well, I think that's a reasonable mark."
Tutor 2	13	75	T2: "Oh really? I wouldn't have given it 92. No, I think that is definitely too high. Mind you, I think I'm probably quite a hard marker. If I was monitoring and it was marked by a tutor and they had given it a mark in the low 80s, I would be fine with that. If they gave it a mark of 85 or above, I would tell them they were being lenient."
Tutor 3	45	85-87	T3: "Well, I would give it a Pass One, I think, yes."
Tutor 4	32	89,90,91	T4: "Yes, I do agree with that, yes, obviously."
Tutor 5	25	85+, late 80s	T5: "Yeah, 92, I suppose if a Pass One is 85 plus, I would have probably upped it a bit to the late 80s but I would have given it a few marks below that…"
Tutor 6	16	87	DB: "Okay, so the essay was given 92." T6: "Okay." DB: "So you agree with that?" T6: "Yes, yeah."
Tutor 7	28	75	T7: "Bloody hell! Really? Sorry, I mean… But I wouldn't have put it above a top end of the Pass Two anyway."

It is important to consider that these comments, made during the interviews, were raw marks that in normal circumstances and to standardise the Open University marking would be balanced with the second marker's decision; and with the third marker's in case of a possible disagreement during the coordination meetings. Therefore, based on this sample size, it is not credible to generalise the result that every tutor marks completely differently and unreliably. As an illustration, consider the following script from the interview with Tutor 1.

> TUTOR 1: "…I suppose I am speaking here as someone who has to support ALs as well in doing this. What we try to do is to have a co-ordination session where everybody talks about what marks they are giving to, you know, we have a debate about, about how we are valuing …"
>
> TUTOR 1: "But those things have to be discussed and there is never, it is inevitable, with the best will in the world that two very experienced tutors can give a very different mark to the same assignment."

Yet, it is significant to note that human marking and assessment may vary depending on several factors whereas automated analysis always provides the same result every time. This supports the argument that there is a benefit to using an automated technology, which could support educators' marking.

The Jaccard analysis results showed a high, significant similarity between Tutor 1 and the XIP highlights. Especially considering that Tutor 1 is a module chair for E000 who looks at the marked scripts and is responsible for guiding the ALs to mark as reliably as possible, holds coordination meetings with ALs to discuss their marking and third marks the essays to adjudicate a mark should two ALs disagree on the mark of an essay, this is promising for further evaluation of the congruency of the XIP's analysis results with the educators' judgement of quality.

Although there is a highly significant similarity between the module chair's, Tutor 1, and the XIP's highlights, in other cases the results were not significant even with the other module chair, Tutor 2. The reason could be dependent on several other factors as discussed above but qualitative data analysis of the rest of this chapter suggests ways in which XIP could be developed in order to yield more significant results. Since the statistical results (Jaccard analysis results) given above did not prove to be reliable, it is important to examine how tutors actually define the attributes of good student writing and how they interpreted what they highlighted. The next section therefore deals with this and describes the qualitative data analysis of the interviews.

6.5 Qualitative analysis of the interviews

After the transcription of the audio records for each interview, qualitative thematic analysis of the interview data was carried out. The responses yielded data for content analysis that permitted theme creation based on the frequency (number of appearances) of responses of the seven interviewees. The themes were derived from the interview data by following the steps of thematic analysis outlined by Attride-Stirling (2001) and Braun and Clarke (2006), and therefore possible themes were not identified before the focus group session. The thematic categories chosen were driven not only by the frequency but also by the research questions asked in this study and the emergence of ideas that are not dealt with by the existing literature regarding quality of writing, and how markers assess writing and disciplinary and level-based differences in writing.

The first section of the interviews, prior to the essay-marking exercise, was about how tutors define quality in student writing, i.e. what elements and key components make a good quality, successful student essay. Based on the thematic analysis of the interview data, five overarching themes emerged: criticality, argumentation with evidence, voice, coherence/orientation/structure, and relevant content, which are explored in more detail in the following section.

6.6 Themes defining good quality student writing

1. **Criticality**

Tutors defined good quality student writing as critical: by showing an ability to interrogate the assumptions that lie behind arguments. All seven interviewees mentioned that the written text should contain analysis of others' arguments rather than just providing a commentary. Students should show the reader that they have carried out critical readings, reviews and provided evidence of that critical engagement with the essay. The following interview scripts illustrate this point.

> TUTOR 1: **"**I would say at this level, Level Three, I would be looking for what I would call a certain amount of criticality: ability to stand back from one's own perspective and to stand back from the perspective of others and evaluate them; to see there are different ways of looking at the same thing, rather than taking anything for granted or taking any – to be examining your own assumptions and the assumptions of others all the time**."**
>
> TUTOR 4: "A really good, a really good piece will be somebody who has tried, and it won't even succeed, but someone who has tried to, to engage with, or, or with the critical readings and with the, the texts, the children's literature itself and put something of themselves in it; so that they've engaged critically with them and tried to put that over in a way that has communicated it reasonably clearly."
>
> TUTOR 6: "The key elements: that they are confident, that they give me confidence as I'm reading that it's not just surface knowledge. That they are not just throwing in key terms thinking they sound good without understanding what the underpinning is. It is not just surface knowledge; they are not just parroting what they've picked up from the blocks. In third level, it is very, very crucial that they are synthesising things."

TUTOR 7: “A good piece of student writing engages with two things I feel: the question, and the material supporting the question. Once they’ve understood the question they then need to be able to critically assess both the primary novels or picture books, and also critically assess the academic supporting essays that are provided to them.”

2. Argumentation with evidence

All seven tutors stated that criticality should be demonstrated through good argumentation in the text, which would be evidenced by the materials, literature and readings that student has reviewed.

TUTOR 2: “Things that we always stress with students are that they must use evidence and argument; in other words, it’s good thinking based on actual evidence, which is usually evidence drawn from the module or drawn from the set text that they have to read.”

TUTOR 3: “But in a good essay I would expect to see direct quotations from the literary text, which the student has analysed, pulled out the significant details of the language and explained their significance for the topic of the essay.”

TUTOR 1: “…key thing about argument is the use of evidence; in the context of this module I would say that is going to involve relevant citation from critical sources, which are provided.”

Tutors also emphasized that students should not only argue based on the supporting evidence but also by criticising contradicting elements of others’ arguments.

TUTOR 3: “I would expect to see an extensive range of research into the relative materials. So that would include research into the background areas of concepts behind the essay topic and also research into the criticism that already exists about that primary text, about that literary work as well.”

TUTOR 7: “rather than replicating, shall we say, Peter Hunt in his essay on Swallows and Amazons says X; and in Swallows and Amazons on page 55, we can

see that this is true. Well that's agreeing and they are backing it up. But I don't necessarily want to see that all the way through. I would like to see occasionally a challenge. Sometimes, of course, the critics will present very strong arguments and it is difficult to disagree with some arguments; they are very strong but I do, I do hope that students will recognise holes and flaws in critics' arguments."

TUTOR 2: "…All right, now give me an example of some of these contradictory elements you have just told me are in there otherwise why should I believe you… So that's a good argument."

3. Voice

Interviewees said that the key element that brings student success is their ability to argue not simply by backing up their arguments with references, but by taking it to another level through the challenge of other peoples' arguments to establish their own argument.

TUTOR 2: "…they are showing you that they have read everything that you wanted them to read; they've really understood it; they've got their own take on it. In other words, they've reached their own understanding and their own thinking. They are not just repeating what they've read so they are actually presenting an argument of their own, which isn't simply agreeing with everything they've read."

TUTOR 1: "And again it's going to involve an ability to stand back from the evidence, and not just to assume that just because it's in print it's true."

TUTOR 5: "So you would have evidence that they had engaged with the module activities, understood the ideas, but gone beyond that and become an independent thinker. So I would hope to see some evidence of something independent and individual in the writing as well as a good knowledge of the course materials."

TUTOR 7: "So I would expect them to read the essays carefully, the ones that are appropriate to their question and also be able to agree or disagree with the critics in

relation to the question. So it takes a great deal of bravery for a student to do that because, I remember from being an undergraduate student myself, that, it seems almost, how can I say, disrespectful to criticise somebody who is so many tiers above you educationally. Someone who has so many letters after their name and who has a Doctor or Professor in front of it and how dare I challenge or disagree with them!"

4. Coherence / Orientation / Structure

All six tutors mentioned that students should guide the reader about the structure of the essay, and signpost what they are going to write about in the beginning.

TUTOR 1: "I think another thing that is clearly important is a sense of coherence. So something that makes you feel as if the parts of the essay fit together somehow in a logical way."

TUTOR 2: "They should tell you at the beginning, I am going to make, I am going to argue this. I am going to make this kind of argument. This is the way I am going to go about it. They should make a statement in the beginning in their introduction about the sort of direction it is going in. …So, so you are kind of – you are orientated at the beginning. That is a kind of orientation at the start and that is really helpful."

TUTOR 3: "I would expect to see a clearly explained line of argument where the student sets out in the introduction what the essay is going to prove and demonstrate and then the rest of the essay presents the evidence for their stance on the question."

5. Relevant Content

All tutors added an important feature that would bring students to success. They put the 'content' theme above the four they mentioned at first. They all emphasised that students should answer the essay question, and what they write should be relevant to

the topic. Even if students successfully achieved the four themes given earlier, if they have not done these within the context and the topic, then they cannot get any marks.

TUTOR 4*:* "I think, I agree argument is important, it is one of the main things but it's no good having a good argument that is not relevant to the question. It is no good having a good argument that doesn't show knowledge and understanding of our module."

TUTOR 6: "A few times I've had some beautiful essays that just haven't answered the question and you end up writing comments that go, 'Had we been discussing…' and you write out the question that they answered. 'Had we been discussing this, you would have scored 90%. Unfortunately, we weren't."

TUTOR 5: "I think I have to say relevance to the question is pretty important and the knowledge and understanding, so probably these two; argument would come third."

TUTOR 2: "If you see a wonderful essay which is about something other than the question that you've set, how do you know that the student is studying that course and how do you know that this isn't an essay from some other course somewhere else, possibly by somebody else? But if it's not answering the question that we've set this year, we've got no evidence that this student is studying on our course really, have we? … There are lots of wonderful essays in the world which are about the set text that we've got in our course but they are not relevant for us because they are not about our course. ((*Laughs*))"

6.7 Other themes

Three other themes emerged during the interviews.

1. **Not everything is equally important (referencing, language)**

   Besides the five key elements that make an essay of good quality, all interviewees added another element but which is not as important, and that is referencing style and language. Tutors expect student writers to acknowledge the cited material; but any typographical errors in doing this do not hugely affect the grade. All seven tutors noted that they mostly neglect these tiny problems in the essay, if they are not getting in the way of the meaning.

   TUTOR 3: "I'm most interested in the quality of the student's ideas and how they've articulated those and, and, you know, and I am possibly slightly less bothered about accuracy of spelling and, you know, totally the accurate formatting and referencing."

   TUTOR 1: "So what I don't like to see when I'm looking at marked essays is a tutor who has treated it like a proof reading or editing exercise and they've corrected every single language error. That is just pointless."

   TUTOR 2: "And then, you know, the final thing, which I think is probably less important; people fuss about it an awful lot but it - enormously about referencing at every level from beginners in their higher education right through to PhD level, people pick on you about your referencing and being accurate. Everybody gets very stressed about it, but it is important, but it's not that important really if you don't put things in italics or you miss something out here and there. … And people can write, oh, clumsily, mess up their references a bit, and so on and still do pretty well because those things are more like superficial technical details."

   TUTOR 4: "I suppose I'd also look for the academic conventions like referencing. I would expect to see that. I wouldn't worry too much if they weren't perfect. I am, I

am really looking for someone who has really put something into the assignment and thought about it."

TUTOR 7: "I do mean grammar, grammar is minor; grammar and spelling and punctuation are minor but they are only minor as long as they don't in any way interfere with me understanding what they are saying. When it becomes problematic for me to understand what they are trying to communicate, that is when I take notice of it."

2. What makes good quality student writing is discipline- and level-specific

All tutors said a good piece of student writing varies between different disciplines.

TUTOR 1: "…the nature of a good piece of academic writing varies and there isn't one standard and I am sceptical of ideas that there is one standard."

Interviewees noted that no matter what the discipline is, clarity is key but the way a student writer argues would differ from discipline to discipline. They especially noted that although they mentioned that criticality and challenging others' arguments are essential features for successful essays, some disciplines do not allow this, and use empirical evidence as truth and fact.

TUTOR 4: "It certainly makes a difference on how it is judged and the grades it gets, yes. I mean I don't, I feel I'm answering your questions with two, with two sort of viewpoints: one as a marker who follows the marking criteria and looks at, you know, the assessment in terms of the marking criteria, and the other is why I am personally impressed by it. I mean I would say clarity for me in whatever, to aim for clarity is important, but I mean I don't mark physics essays but obviously I've read that sort of thing in the past so I know it is obviously different in the way that it tends to be more objective."

TUTOR 7: "when I was studying to be a librarian, … that was within the social science discipline; so I've come from social science into literature and I do find

myself having to drop certain pre-conceptions of students on this course with my previous discipline of social science. And I've found, in social science, there was almost no encouragement to challenge academic theory or critical explanations of things, and it was really to use your critical sources really to produce, I suppose a discursive piece rather than a critical piece or an analytical piece."

TUTOR 5: "Well, particularly in relation to E000, students coming from literature have been taught to value an elegant writing style and they are sometimes more discursive; so with discussing something at length in some depth. Whereas somebody from linguistics might be more focussed on the empirical evidence and as you say a clear style is very important but it doesn't maybe need to be such a literary style, in fact probably they would not want it to be a literary style; it is almost slightly more of a scientific style. And they would, yeah, they would want to back up every claim they make with some kind of empirical evidence; whereas with literature students it is more possible to be a bit speculative and to write about ideas without necessarily tying every single one to empirical evidence."

Interviewees also noted that the expectations of student writers vary at different levels. Tutors expect their students to be more critical, and analyse and synthesise the materials well and construct arguments at Level 3.

TUTOR 5: "So the Level One questions might be focussing more on a particular topic, so you pull together the knowledge and information you have on that topic whereas at Level Three you would expect to do more synthesis of ideas from different areas and bringing them together to construct an argument or to construct an explanation. And I would expect more critical engagement with academic writing in Level Three."

TUTOR 7: "A Level One student – particularly in this university, I would expect to have limited skills and would typically, be perhaps along the – on our grid along the Pass Three/Pass Four levels and be like a bird learning to fly. …I would expect

to see far less of what I like to see in terms of relevance and ability to engage with and criticise, and would probably be writing more commentary and discussion rather than analysis, because analysis to me is a skill that is developed rather than innate in any person."

TUTOR 4: "I would expect more, more of it to be done more successfully in terms of – for a higher level. I mean something that would, so something that would perhaps just pass at Level Three would probably get a higher grade at Levels One and Two. Yeah, and it – a sort of beginner level, well not beginner but at a sort of introductory level I wouldn't really expect the academic conventions to be there. I wouldn't expect them so much but I would expect them to be there by Level Two."

3. Tutors experience some problems with marking

During the interviews, tutors talked about some problems with marking that they experience. One of the problems they raised was about the subjectivity of human marking.

TUTOR 6: "We already get tutor-marking notes and, the trouble is it is a very subjective thing is marking. I think if they were going to try and get us all – if, if the OU wanted to make marking more of a science and less objective, I think what would happen, what would be helpful is if we had maybe a meeting after everybody marks the first EMA to actually moderate it. … -you know, have this discussion as to, oh, I think that's a bit high for that one, who was marking it and why did they give it -? You know, could I be persuaded? Did I miss something?"

Tutors mentioned they only have a limited time to mark an essay, hence they spend too much time marking papers and feel pressured with hours of grading.

TUTOR 7: "… we only have 45 minutes to mark these you know."

DB: "Really, do you set only 45 minutes for marking -?"

TUTOR 7: “That is the agreed amount of time we have for downloading, reading, annotating the script and doing the comment for, and giving a grade and sending it back – that is what we are contractually paid for. Now I find it takes at least an hour. So most tutors you ask are, now this is getting very political, are underpaid for marking because people severely underestimate.”

TUTOR 2: “… I feel that there is just too much. So fewer assignments to mark and the amount of preparation in marking an assignment like this, even though I am familiar with it, is mega. I would have to read all those readings again.”

TUTOR 2: “People who are marking EMAs, they have 60 scripts, sometimes 80 scripts to mark in a matter of a couple of weeks and then individuals. I think it is really hard to really see the value in an essay when you are under that time pressure.”

Additionally, markers mentioned that they struggle with giving feedback, commenting and annotating students’ essays, which is even more time consuming to make sure they gave a clearer explanation about why they have given a specific feedback to their students.

TUTOR 7: “How do you tell that student in writing really, in typing it to them, how do you put it over to them in a way that means something to them? That’s the real problem and that’s time consuming too, very, very time consuming. That is why this was marked in 45 minutes and it would be double that if it was to a student.”

TUTOR 3: “Really, this only applies to TMAs, but the annotation, actually the comments; that’s the bit I hate the most so I always do that first because it is so time-consuming.”

TUTOR 5: “if you say to a student, you know, you need to be more critical – they can’t just do it because they don’t know what you mean. If they could just like that they would have done it. (…) So you need to show them, give them an example. And you – and the example needs to be tied quite closely to the student’s own work

because that is where their subject was. And I think that is something again that tutors do find quite hard to do."

4. **Tutors feel uncomfortable discussing how technology might be helpful to overcome their problems**

When tutors raised the problems they experience with assessment, they were asked whether they would consider using a computational language technology that might potentially help them to overcome such problems. They stated the worry that they might be replaced by technology.

> TUTOR 7: "Would that be automatically done with a computer?"
>
> DB: "Yes."
>
> TUTOR 7: "Wow, weird, I'm beginning to feel I'm replaced by technology now. But you don't need me at all."
>
> DB: "No, the technology tries to help you."
>
> TUTOR 6: "Okay, it could be but again just because you've used a phrase doesn't necessarily mean you've used it correctly or in the right context (…) I'd feel uncomfortable using that because I don't know, it's a step too far I think. I know this is probably the brave new world and I should be embracing new technology and new ideas but it makes me worry…"
>
> TUTOR 5: "No, no it wouldn't help me any more."
>
> DB: "Why?"
>
> TUTOR 5: "Because I know what I'm doing. Whereas five years ago, ten years ago I didn't so it might have then but I don't need it now. So I don't mean that to sound arrogant, you know, I'm confident in my ability to evaluate the students I face."

6.8 Discussion

This study provides a systematic investigation of what human markers value prior to adapting the automated technology, XIP. In the literature review, it was emphasised that, although there are some guidelines and marking criteria that can be used to understand what markers value, there is lack of understanding of what human markers actually do in their evaluation (Attali, 2013). Therefore, this study has supported understanding of the markers, what they really value in student writing. One-to-one interviews provided more information than the available marking rubrics, about how academic tutors mark, and what they specifically give credit to. Additionally, in line with the literature, the problems of subjectivity in human marking, time limitations, the need to provide better feedback and examples, and the suspicion and tension towards the use of automated text analysis became evident in the one-to-one interviews with tutors.

One of the primary goals of automated essay analysis is to ensure that human markers think similarly about what constitutes high or low quality student writing so that the machine can be used effectively to analyse the core, essential criteria. An ideal would be that it takes out the subjective elements that human markers might tend to include. Although the quantitative analysis of the essay highlighting exercise did not provide similar results, the qualitative analysis of the interview data provided the key elements that make for good quality successful student writing in this context.

From the one-to-one interviews with markers, it is noted that good quality of student writing cannot be defined generically; therefore, there is not a gold standard for student writing. Good quality of student writing is discipline- and level-specific. This triangulates with the results of the quantitative analysis explained in Chapter 5. In the XIP analysis of student writing, it was observed that the current form of XIP did not work well for all disciplines and it produced different results for different levels.

Interviewees noted that they expect to see an improvement in student writing at higher levels. Students are expected to write more discursively, providing commentary, at lower levels, simply to provide information and show understanding of the course materials. At higher levels, students need to analyse knowledge, synthesise information, criticise and argue, so they demonstrate that they have become critical thinkers and writers. This triangulates with the findings discussed in the previous chapter.

The analysis of the BAWE corpus showed that the XIP categories like SUMMARY, where the writer should summarise the goals or the results of the article, are found in student writing at Level 1 or 2. On the other hand, the categories that require higher-order writing skills such as OPEN QUESTIONS, where the writer describes problems which have not been solved, are found at Level 3.

From the interviews with the markers, five key elements emerged that make for good quality student writing and which are related to the existing XIP categories below:

1. **Criticality**

   Interviewees defined successful student writers as critical thinkers who read the relevant material and critically analyse its arguments. Considering the XIP categories, this could be captured by several categories, including OPEN QUESTIONS, describing problems that have not been solved, and EMPHASIS, emphasising the importance of ideas. Additionally, the category of CONTRAST can capture the criticality of the students; however, tutors mentioned one different aspect that is not directly CONTRAST but could be assumed to be a subset to this, which is 'SYNTHESIS'. In order to contrast two or more research ideas, student writers should read other researchers' arguments, synthesise them and then contrast these either with their own ideas or with each other. Currently, in XIP, there is no category that shows the sentences conveying the rhetorical meaning as 'SYNTHESIS'; which could be added as a new category.

2. **Argumentation with evidence**

   Successful student writers are expected to argue based on supporting or contradictory evidence. Considering the XIP categories, supporting evidence could be captured through the XIP category BACKGROUND, describing background knowledge necessary for understanding the article's contribution. Contradictory evidence could be captured by the XIP category CONTRAST, describing tensions, contrasts between ideas, models or research directions.

3. **Voice**

   Markers said that in good quality student writing, students act as independent thinkers. Readers would hear the voice of the student writer who constructs new arguments based on the relevant literature reviewed. Similarly to the two themes BACKGROUND, describing background knowledge necessary for understanding the article's contribution, and CONTRAST, describing tensions, contrasts between ideas, models or research directions, discussed above, some of the XIP categories could possibly capture this when they come together. For example, the student writer would provide the existing arguments through the category BACKGROUND, criticise these with CONTRAST, and then express what has not been mentioned through OPEN QUESTIONS (describing problems that have not been solved), providing their own approach through NOVELTY (conveying that an idea is new).

4. **Coherence / Orientation / Structure**

   The markers noted that in good quality student writing, the writer would guide the reader through the structure of the essay, and signpost what the essay would be about at the beginning. Although XIP currently returns no results about the structure or coherence, one of its categories can capture how the writer sets out the

arguments and the aim of the essay which is the category SUMMARY, where the writer summarises the goals or the results of the article.

5. **Relevant Content**

   Markers place the content above all other features. However, considering the fact that machines cannot read, understand an essay and interpret its meaning, XIP would not be able to tell how relevant the written material is to the essay question or the topic. None of the XIP categories could capture this. However, the focus here is on ways to support the educators in their essay assessment processes, not to automate the marking. Since the markers are able to decide whether the content is relevant or not, this does not create any problem with the idea of using XIP. XIP would not be the judge here, it would only be the lawyer that helps the jury judge the final decision.

Although it seems that some of the XIP categories could be tied well with these five elements, the quantitative results of E000, as discussed in Chapter 5, suggest that not all the XIP categories, in its current form, could capture these key elements. Many of these suggested categories did not emerge as significant in the previous chapter. The possible reason for this is that the XIP categories have been established by the literary styles used in experienced researchers' journal writing. Although at Level 3 student writers are expected to write critically as experienced researchers would, their style is unlikely to be as good as or similar to that of an experienced researcher's.

The significant XIP categories, identified in the previous chapter, that increased the essay grade positively for E000 were BACKGROUND, and CONTRAST. This shows that the five key elements of student success do not easily relate to the existing XIP categories above. However, the XIP categories can be tuned for specific requirements; new categories could be created with new rules that could capture these key elements. It is possible to tailor new categories to capture these elements discussed above. However, doing this

requires not only knowing what markers value in good quality writing, but also requires discussion of the current form of the XIP with them. The next chapter reports on a focus group study that enabled educators to discuss the XIP analysis, highlights, and categories.

7

FOCUS GROUP

7.1 Introduction

This chapter addresses the third subsidiary research question: "How congruent is the XIP analysis of student essays with educators' judgement of quality?" Answering this question required an investigation into how educators define the quality of student essays and to what extent the XIP analysis can capture this.

The multiple regression study, explained in Chapter 5, suggested that promising results could be obtained from relating XIP analysis to undergraduate essay marks, despite the fact that the tool had not been developed for this particular purpose and context. Since it is important that academics and educators trust a new tool that analyses writing, it is important to hear what queries or even doubts they might have about how such a tool works, how similar its output is with respect to their own judgement of quality, and how it could be improved.

Although one-to-one interviews with educators, as explained in Chapter 6, illustrated to some extent how educators assess student writing and how they define the quality of that writing, this was restricted to individual opinion. This meant there was no room to explore different perspectives of how quality might be judged.

This research question requires the evaluation and comparison of multiple perspectives about exploring the possibilities of integrating the XIP tool into educational contexts in order to analyse undergraduate writing, which was explored by using a focus group. The

next section of this chapter reports on the design details of the study and how the data were collected; this is then followed by the explanations of the participants. An account of how the data were transcribed and analysed is then given, and finally the findings are reported.

7.2 Study

The preferred number of carefully selected similar types of focus group participants is six to eight (Krueger & Casey, 2000) with a maximum ten people suggested per session (Morgan, 1996). Having a maximum six to ten people in each focus group session ensures that all members of the group will have enough time to speak and share their opinions during discussions. One or preferably more sessions should be conducted (Morgan, 1996). This number depends on the overall research design; more sessions are required especially if the focus group is the only technique used in the project. Although focus groups can be employed as a self-contained method, they can also be used in combination with other research methods mainly with individual in-depth interviews (Morgan, 1996). As a focus group is not the only technique used in this thesis, it has been employed as a further study to triangulate the data collected from individual interviews. In fact, one focus group discussion was held with six participants at the Open University on 2nd December 2015. The study was advertised in the Faculty of Education and Languages (FELS) since many academics in this faculty have a particular interest in the area of student academic writing and have experience in teaching and marking student essays. An invitation was sent out to a research group called the "Academic and Professional Literacies Forum" (AcLit). This group was targeted not only because these forum members, who are writing teachers and researchers, meet monthly to discuss topics around student academic writing but also because this research topic had been presented to the group two years before the focus group study, and participants at that time showed great interest and provided supportive criticism of the studies described in this thesis. The participant selection was not confined to these people, however, and an advertisement was also sent out to all those in the FELS

department who carry out research in student writing and who matched the criteria of experience in teaching and marking. Due to retirements, maternity leave and several other reasons, two out of six participants were from the previous forum group presentation held two years earlier, and the rest were new to the topic. These participants were different from the educators who were interviewed before. Therefore, the focus group discussion was started with a presentation on how the XIP works, what research has been carried out so far and its results. This allowed participants:

- to learn about the tool itself, how it has been developed and for what purpose
- to ask questions about its basis in linguistic theory
- and to see the promising quantitative study results (as explained in Chapter 5) that allowed them to understand the reason for conducting the focus group study.

The focus group discussion was designed to observe the group interaction and to gather information from educators with respect to the following outcomes:

1. to understand what participants think about the accuracy of the XIP results
2. to understand what participants like and do not like about the tool
3. to understand perceptions of the applicability of the XIP in an educational context
4. to capture participants' perceptions of how the XIP tool can be integrated and adapted successfully in an educational context
5. to discuss possible actions to make the XIP tool better for the purpose of using it in an educational context.

All focus group participants were present for the presentation as well as for the focus group study itself. A question-and-answer session followed the presentation and then, after a break, the focus group study began. The study was in three parts. Before the first part began, participants wrote down their initial thoughts and/or misgivings about the possibility of using the XIP tool in education. Participants were then given two pages of student writing and, without guidance, they were asked to highlight the sentences that they

thought illustrated good quality writing, with respect to good critical, argumentative or analytical statements. In the one-to-one interviews, that activity was followed by questions regarding assessment techniques and styles; but this time, since they had already been told about the XIP tool, the XIP analysis results were shown. After receiving their reactions to the XIP results, participants were then asked to discuss the sentence category (i.e. Summary, Background, Contrast) that XIP might have assigned to each salient sentence and whether they agreed with XIP's choice.

In the second part, after participants were informed about the potential of XIP, they were asked whether they would consider using XIP to analyse their own students' writing if the XIP tool were to be made publicly available to analyse any form of student writing. They wrote down three important features that would convince them to use the tool. After participants had shared their ideas, they discussed what would be the most important factor in their decisions. In the final part, participants discussed what might need improvement and what sorts of change they would make to improve the system. At the close of the session, participants were asked to write down their final thoughts and/or doubts about the XIP.

The next section describes the analysis of the first part, the highlighting activity.

7.3 Quantitative analysis of the highlighting activity

The first part of the focus group session was the highlighting activity, where educators highlighted the sentences that they thought illustrated good quality writing, with respect to good critical, argumentative or analytical statements. The objective of this activity was to get educators to understand the XIP tool, conduct a follow up discussion regarding what they think of the XIP highlights for the same piece of text, and to explore reliability issues with human marking, as raised in the previous chapter.

The highlighting activity data had also been used for quantitative analysis to examine any overlaps between participants' and XIP's highlights. Table 7.1 shows for each sentence

whether participants or the XIP highlighted that sentence. Participants are shown as P1, P2, P3, P4, P5, and P6. The sign 'x' shows if the sentence was highlighted by individual participants. Similarly, the XIP column, shaded in yellow, shows the sentences picked up by the XIP tool. There were 13 sentences in total. The XIP tool identified nine of these sentences as salient. The reason why the P5 column is shaded is because that this participant did not highlight the sentences fully, although she had been asked to do so. As this participant highlighted the whole text in phrase level for almost every sentence, her input was discarded for this part of the analysis.

Table 7.1 Sentences highlighted by focus group participants and by XIP

Sentence	P1	P2	P3	P4	P5	P6	XIP
1) While the history of literature written specifically for children is relatively short, children's literature has proved to be a rapidly evolving genre, subject to a wealth of critical debate.	x	x	x	x		x	**x**
2) At the heart of this debate lies the challenge of defining the purpose of children's literature and in particular, the question of whether these texts are largely didactic, with the agenda of conveying ideas and knowledge, or whether their focus remains primarily on entertaining their young readers.	x	x	x			x	**x**
3) Peter Hunt highlights how children are 'susceptible to new ideas' (Hunt, 2009a, p.15), and that consequently their literature has the potential to make a lasting impression to the extent that 'it is inconceivable that these texts have not shaped society in fundamental and lasting ways' (Hunt, 2009a, p.15).	x	x	x			x	
4) In addition, Kimberley Reynolds argues in favour of the radical potential of children's literature, claiming that it encourages the development of the 'tools necessary for thinking about themselves and the world they inhabit' (Reynolds, 2009, p.99), and as a breeding ground for new ideas, has the power to instigate social change.	x	x	x	x		x	**x**
5) Both critics, therefore, suggest that a didactic element exists in children's literature and by extrapolation, that literature without agenda cannot exist.	x	x	x	x		x	**x**
6) However, a convincing case for a child's ability to ignore all but the most transparent of agendas has been contended, with children being argued to possess 'the happy faculty of discovering and assimilating that which suits him, and passing by the rest' (Field, as cited in Dusinberre, 1999, p.44).	x	x	x	x		x	**x**

7)	This prompts the question: can an agenda be argued to exist in a text if it remains undetected by a proportion of its readers?	x	x	x	x		x	
8)	In order to approach the question of whether an agenda of some kind can be discerned in all children's literature, this essay will examine the incidence of perceived overt and hidden agendas in Roll of Thunder, Hear My Cry (Taylor, 1976), Little Women (Alcott, 1869) and Peter Pan (Barrie, 1904).	x						**x**
9)	All three texts convey aspects of children 'growing up', however this critique will highlight that the definition of their agendas is dependent on their readership.	x	x	x	x		x	**x**
10)	To illustrate this thesis, the widely-accepted overtly didactic objectives of Roll of Thunder, Hear My Cry will be scrutinised and the debate over the nature of its agenda of communicating racial prejudice will be examined to demonstrate how 'overt' agendas are still debatable.							**x**
11)	In addition, to exemplify how perceived agendas may differ among modern readers, the debate over the feminist agendas in Little Women will be outlined.			x				
12)	Finally, Peter Pan will be scrutinised to illustrate how the subjectivity of both the overt agenda of 'growing up well' and the covert agenda of 'suppression of adult heterosexuality' (Rose, 2009) is enhanced when the drama is adapted for film.			x				
13)	This essay will utilise these arguments to assert that since we cannot 'make assumptions about what any reader but ourselves perceives' (Hunt, 2009a, p.17), the agenda of a children's literature text is subjective and thus while all children's literature might be argued to contain an agenda of some form, this will vary from reader to reader.							**x**

In order to explore the overlap between participants' and XIP's highlights, the similarity of these outputs was calculated. Having XIP's and the participants' highlights, the similarity between participant pairings, as well as between XIP and each participant was measured by using a Jaccard similarity index (see section 6.3 for calculation of Jaccard similarity index).

Jaccard analysis was performed between participants to find out whether they agreed with each other and if their highlights were similar. The following matrix, table 7.2, shows the common number of sentences highlighted by participants. These numbers are used in the calculation of the Jaccard similarity coefficient, as given in table 7.3 below.

Table 7.2 The common number of sentences highlighted by participants

	P1	P2	P3	P4	P6
P1	-	8	8	6	8
P2	8	-	8	6	8
P3	8	8	-	6	8
P4	6	6	6	-	6
P6	8	8	8	6	-

Table 7.3 Jaccard Similarity Index matrix for each participant pairing

Jaccard Similarity Index coefficient	P1	P2	P3	P4	P6
P1	-	0.88	0.72	0.66	0.88
P2	8/(9+8-8) 0.88	-	0.80	0.75	1
P3	8/(9+10-8) 0.72	8/(8+10-8) 0.80	-	0.60	0.80
P4	6/(9+6-6) 0.66	6/(8+6-6) 0.75	6/(10+6-6) 0.60	-	0.75
P6	8/(9+8-8) 0.88	8/(8+8-8) 1	8/(10+8-8) 0.80	6/(6+8-6) 0.75	-

The significance test results between each participant pairing is given in next table, table 7.4. The required J values were obtained from Real's (1999, p.31) table 1 based on the associated N value.

Table 7.4 Pairwise significance test results between all participants

N,J, required J & significance test	P1	P2	P3	P4	P6

P1	-	N=9, J= 0.88 Required J value >= 0.6667 **Significant**	N=11, J= 0.72 Required J value >= 0.6364 **Significant**	N=9, J= 0.66 Required J value >= 0.6667 **Not significant (very close)**	N=9, J= 0.88 Required J value >= 0.6667 **Significant**
P2	N=9, J= 0.88 Required J value >= 0.6667 **Significant**	-	N=10, J= 0.80 Required J value >= 0.70 **Significant**	N=8, J= 0.75 Required J value >= 0.75 **Significant**	N=8, J= 1 Required J value >= 0.75 **Significant**
P3	N=11, J= 0.72 Required J value >= 0.6364 **Significant**	N=10, J= 0.80 Required J value >= 0.70 **Significant**	-	N=10, J= 0.60 Required J value >= 0.70 **Not significant**	N=10, J= 0.80 Required J value >= 0.70 **Significant**
P4	N=9, J= 0.66 Required J value >= 0.6667 **Not significant (very close)**	N=8, J= 0.75 Required J value >= 0.75 **Significant**	N=10, J= 0.60 Required J value >= 0.70 **Not significant**	-	N=8, J= 0.75 Required J value >= 0.75 **Significant**
P6	N=9, J= 0.88 Required J value >= 0.6667 **Significant**	N=8, J= 1 Required J value >= 0.75 **Significant**	N=10, J= 0.80 Required J value >= 0.70 **Significant**	N=8, J= 0.75 Required J value >= 0.75 **Significant**	-

The Jaccard analysis results showed a significant similarity between the highlights of most participants, who agreed with each other except P3 and P4.

Jaccard analysis was also performed between participants and the XIP. The following table, table 7.5, shows the total number of sentences highlighted by the participants and the number of common highlights between the XIP and each participant.

Table 7.5 Total numbers of sentences highlighted by each participant and common sentences with the XIP

	Total number of sentences highlighted by participant	Total number of common sentences highlighted by XIP and participant
P1	9	7
P2	8	6
P3	10	6

P4	6	5
P5		
P6	8	6

Jaccard similarity index results for all possible pairings are given in the following table, table 7.6, with calculations. The significance test results between XIP and each participant is also shown in this table. Required J values were obtained from Real's (1999, p.31) table 1 based on the associated N value.

Table 7.6 Jaccard similarity index for each participant and XIP

Participant & XIP pairings	Jaccard Similarity Index	N	Required J	Significance test
P1 & XIP	7/(9+9-7) = **0.6363**	11	>=0.6364	Not significant (very close)
P2 & XIP	6/(8+9-6) = **0.55**	11	>=0. 6364	Not significant
P3 & XIP	6/(10+9-6) = **0.46**	13	>=0. 6154	Not significant
P4 & XIP	5/(6+9-5) = **0.50**	10	>=0. 70	Not significant
P6 & XIP	6/(8+9-6) = **0.55**	11	>=0. 6364	Not significant

The Jaccard analysis results showed no significant similarity between the highlights of XIP and the participants.

7.3.1 Discussion of the quantitative analysis results

The similarity between participants was found to be significant except in two cases as shown in table 5. The pairwise comparison results showed that most of the participants agree with each other, although not P3 and P4. Considering the Jaccard similarity results, given in the previous chapter (chapter 6: one-to-one interviews with markers), there were no significant similarities between any of the tutor pairings. Therefore, the original thought was that perhaps the educators' marking would be significantly different in this study as well, but this was not the case, the participants largely agreed. The reason could be that in

the previous study all the interviewed human markers were tutors of the same module, an essay from which had been used for the study. Therefore, they were looking at the content and considering whether what has been argued by the student writer was accurate. Additionally, human markers used their own marking guidelines in the previous study. Therefore, they looked not only at the argumentative sentences in writing but also at several other aspects like the accuracy of the content, coherency, referencing styles etc. In the focus group, however, educators were not familiar with the content of the given text and they were asked to focus only on the sentences that they thought illustrated good quality writing, with respect to good, critical, argumentative or analytical statements. The similarity between the highlights of XIP and each participant were not found to be significant. The reason for this could be that there were thirteen sentences to work with in total. In the Jaccard table, N is the number of sentences that are highlighted by at least one of the two participants whose highlights are compared, A+B-C. There can be a lot of overlap, however, between participants (or the participant and the tool) just by chance. To illustrate this point, consider the following examples:

- in a situation where XIP picked up five and the participant highlighted six sentences where four sentences were the same with XIP: A = 5, B = 6, C = 4, and the Jaccard index is 4/(5+6-4) which is 4/7 or 0.571. N is seven, and J is 0.571, but to get a significantly high correlation with $p < 0.05$, a J index of 0.7143 or bigger is needed. This would be reached if the participant had highlighted seven sentences including all five that XIP highlighted.
- if the participant highlighted eight sentences including all five that the tool had highlighted, then A = 5, B = 8, C = 5, N = A+B-C = 8, and the index is 5/8 or 0.625, but with N = 8 you would need an index of 0.75 to be significant at 0.05, and it is not possible to get such a large index with five highlighted by the XIP.
- as another example, if the participant highlighted four of the tool's five; but did not highlight any that the tool did not highlight, then A = 5, B = 4, C = 4, N = A+B-

C=5, and the index is 4/5 = 0.8, and in the row for N = 5, that is significant at $p<0.05$.

Therefore, it cannot be concluded that XIP and participant highlights are not similar. The highlighting exercise was carried out in order to help educators understand the XIP tool and to explore issues with marking. Therefore, with such a small sample it was unlikely that statistically reliable or valid results would be obtained and, indeed, the Jaccard analysis showed no significant patterns.

Likewise, although the results show that there was a significant similarity between tutors, (except in two cases), considering the chance factor, and the focus group environment in which participants could see their peers' highlighting work, could talk to each other and look at what their colleagues were doing, the results cannot be accepted as fully correct. It should be acknowledged, however, that this issue was dealt with carefully during the individual interviews, as given in Chapter 6, in which tutors solely highlighted the sentences from a full length student essay, from a larger and reliable sample size, with no discussion with anyone else.

Additionally, although the statistical results showed that the participants' selections were significantly similar, the following table, table 7.7, demonstrates the discrepancy between participants' decisions about the labels given for each XIP highlight. They had different ideas about the interpretation of the rhetorical function and meaning of students' sentences. Although it is true that there was some level of consensus between the participants' selections, their thinking was not always the same. Their decisions were not always about the same sentences, but they disagreed with each other to the same extent as they did with XIP. Obviously, participants had not received training about XIP's coding system, and therefore they could not use such a coding system; but when they asked, there was never any agreement in the first instance. They discussed their decisions for about three to four minutes which then subsequently changed their decisions in some cases.

Table 7.7 The labels of each participant for the XIP highlights.

C=Contrast, B=Background, S=Summary, E=Emphasis, N=Novelty
-=if participant did not produce any information
Salient = If the XIP did not label the sentence with any rhetorical type but simply highlighted
Grey shaded areas are skipped during the session due to time limitations.

Sentence	P1	P2	P3	P4	P5	P6	XIP
While the history of literature written specifically for children is relatively short, children's literature has proved to be a rapidly evolving genre, subject to a wealth of critical debate.	C	C	B	B	C |B	B	**E | C**
At the heart of this debate lies the challenge of defining the purpose of children's literature and in particular, the question of whether these texts are largely didactic, with the agenda of conveying ideas and knowledge, or whether their focus remains primarily on entertaining their young readers.	B |C |S	S	B	B |C |S	C	B |S	**C**
In addition, Kimberley Reynolds argues in favour of the radical potential of children's literature, claiming that it encourages the development of the 'tools necessary for thinking about themselves and the world they inhabit' (Reynolds, 2009, p.99), and as a breeding ground for new ideas, has the power to instigate social change.	B	E	S	B| S	B	S	**N**
Both critics, therefore, suggest that a didactic element exists in children's literature and by extrapolation, that literature without agenda cannot exist.	B	S	B	E	C	E	**C**
However, a convincing case for a child's ability to ignore all but the most transparent of agendas has been contended, with children being argued to possess 'the happy faculty of discovering and assimilating that which suits him, and passing by the rest' (Field, as cited in Dusinberre, 1999, p.44).	C	C	B |C	B	-	C	**C**
All three texts convey aspects of children 'growing up', however this critique will highlight that the definition of their agendas is dependent on their readership.	B |C	C	C	S	B	S	**Salient**
This essay will utilise these arguments to assert that since we cannot 'make assumptions about what any reader but ourselves perceives' (Hunt, 2009a, p.17), the agenda of a children's literature text is subjective and thus while all children's literature might be argued to contain an agenda of some form, this will vary from reader to reader.							**S | C**
In order to approach the question of whether an agenda of some kind can be discerned in all children's literature, this essay will examine the incidence of perceived overt and hidden agendas in Roll of Thunder, Hear My Cry (Taylor, 1976), Little Women (Alcott, 1869) and Peter Pan (Barrie, 1904).							**S**
To illustrate this thesis, the widely-accepted overtly didactic objectives of Roll of Thunder, Hear My Cry will be scrutinised and the debate over the nature of its agenda of communicating racial prejudice will be examined to demonstrate how 'overt' agendas are still debatable.							**Salient**

Having various labels which did not match most of the time, neither between the participants nor with the XIP could either be explained by the fact that XIP's labelling system does not work, or that the participants needed extra training before using such a coding system. Considering the conclusions raised in the previous chapter (chapter 6: one to one interviews with markers), XIP needs additional categories such as SYNTHESIS and it needs to unpack the underlying meaning of each category. It can also be interpreted to mean that the current naming of XIP might be adapted for educators' language. Further dialogue with its users is necessary before the system can be for clarification.

The highlighting exercise was carried out in order to help educators understand the XIP tool and to initiate follow up discussions for qualitative analysis, which is explained in the next sections. It was not expected to yield valid quantitative results with such a small sample of text containing 13 sentences, and the Jaccard analysis confirmed that no valid results were available; but it did provide valuable information on how educators think differently to each other, and what they value in student writing differs, as concluded in the previous chapter as well.

The next sections describe the transcription of the focus group data, the qualitative analysis of the second part of the session, and the discussion of its results.

7.4 Transcription

The focus group session was both audio and video recorded, using high-quality playback equipment. A full verbatim transcription (Poland, 1995) approach was followed. This approach involves noting down both the non-verbal actions like gestures, mimics, gazes, nods; and verbal actions signifying hesitations, ignorance, laughs, sarcasm, confusion, excitement, like confusion in the tone of the voice, murmurs, hums, okays, etc. Adding observational data like facial expressions made it possible to to observe how people's ideas

had changed and were also influenced by others. To illustrate this point, the following example transcript is reported:

"D.B: …without any linguistic theory behind it.

PARTICIPANT 1: Aha! ((*eyebrows up*))

PARTICIPANTS 3 & 4: ((*cynical smiling, looking at each other, nodding*))"

As seen in this example script, participants' verbal and physical actions signify their disapproval of the fact that there is no linguistic theory behind XIP. This approach has been consistently applied to the transcription. To achieve this, the initial transcription was made using the audio playback, and data about gaze and gesture, verbatim data, were added later after following the video recordings. A summary of the conventions used in transcriptions and guidelines relating to the format are given in the table provided in the APPENDIX G.

After transcription, qualitative thematic analysis of the qualitative focus group data was carried out. The responses yielded data for content analysis that permitted theme creation based on the frequency (number of appearances) and intensity (emotion) of the responses of the six participants. The themes were derived from the interview data by following the thematic analysis steps outlined by Attride-Stirling, 2001; Braun and Clarke, 2006, and therefore possible themes were not identified before the focus group session.

The themes came out especially where there were general agreements/disagreements and inconsistencies about a topic. "A theme captures something important about the data in relation to the research question" (Braun and Clarke, 2006, p.10). Therefore, the theme categories were driven not only by the frequency and intensity, but also by the research questions asked in this thesis and the emergence of ideas that are not dealt with by the existing literature.

All the sections including the presentation before the study were transcribed following the approach detailed above. Another reason for transcribing the presentation was to capture the initial reactions of the participants towards the research results and the whole concept

and idea of automated analysis of text and its use in higher education. This also helped to confirm and triangulate the notion of some of the arguments presented in this thesis.

7.5 Analysis of the transcribed data

Analysis was undertaken both of the verbal data and the observational data of facial expressions. Although it is possible to carry out quantitative analysis of qualitative focus group data (Stewart et. al, 2007), numerical analysis of focus group data is not a preferred option (Grudens-Schuck et. al, 2004) because it may mislead a reader by giving the impression that results by percentage present major findings that are true for a much wider population, which is not the case. Focus groups rely upon words and reactions of the participants. Therefore, the analysis here focuses on patterns formed by words, reactions, themes or perspectives.

When analysing the data, the focus was both on the individual (i.e. why one person liked the idea and why the other one did not) and the group, especially when there was a particular general agreement, together with how a group consensus was reached or not.

7.6 Findings

From the qualitative analysis of the focus group data, five overarching themes emerged, which are explored below.

Theme 1: Quality

During the focus group session, participants talked about what makes good quality student writing. They acknowledged that some students write exceptionally and move beyond what is common in student writing. In the given example, they defined the student writer as a skilful dense writer, meaning s/he used a range of wordings which are relatively infrequent but certainly show that s/he was a thoughtful writer.

PARTICIPANT 5: "…And I think what this writer is doing is quite a lot of exceptionality, you know. I mean there are some fantastic things like, 'has the potential to', 'the question of whether', 'a breeding ground for', 'by extrapolation', 'a convincing case has been contended', 'widely accepted', 'debate over', 'exemplify how'…"

((General Agreement))

PARTICIPANT 1: "So in other words they are using a whole range of things that are perhaps relatively infrequent, but certainly their density makes this a very good writer…"

Participants pointed to aspects of form and beauty as being a sign of quality. Their points about subtlety, beauty and flow demonstrated their views about what good student writing is. They also emphasised, however, that these are not necessarily taken into account in marking rubrics as a sign of quality. Even though markers need to work on the margins of the rubric, they would recognise and credit the beauty and exceptionality, as almost bordering on one of the criteria, especially in the field of arts.

PARTICIPANT 5: "…I would say even one of those 'a breeding ground for' you'd think: this is a thoughtful writer. So frequently what you are measuring inevitably is frequency, which is conformity, or you know, common or garden – not exceptionality at all."

PARTICIPANT 5: "And I mean that's inevitable; how do you measure exceptionality?"

PARTICIPANT 7: "Yes, because you can't very well have 'breeding ground' as one of your things that is – and yet, and yet in situ, in this context it is a skilful-"

PARTICIPANT 5: "Some of those argue in favour, and potential – there are some that will be borderline quite common, good, very good and quite common in academic writing. But there will be some here that will be really special."

PARTICIPANT 3: “I think it’s interesting, really interesting – so closely related to the Locust and the Michelangelo – we had these slides on the wall today in the PACE session and one of them is a slide with a diagram of a locust and the other one is a diagram of, is a photograph of Michelangelo’s David. I was trying to talk about structure – and so yes, your locust, your exoskeleton, all of the joins are there – all of the bits of meta-language are there and that might be the way you need to write in particular situations.

But then there’s that kind of higher reality, which is the structure that is there but it’s not like, it’s subtle; it’s – it’s got kind of subtlety and flow as well. I didn’t mention the six-pack but the kind of – the idea that you are looking at something of beauty because it’s highly-structured but it’s not – there’s something organic about it as well.

And I suppose in a way what we’re talking about is – are those organic bits and perhaps that’s why there are some of those essays that are really outstandingly good but don’t have any of this language in them, because that person is like – they’ve left, they’ve left this convention behind and they are sort of -.”

[BREAK]

PARTICIPANT 3: “I was just going to say that would be a feature of the arts probably more than the sciences but having seen the film about Stephen Hawking he broke all the rules in his thesis and so, you know, even scientists can recognise the exceptional but certainly in the arts I think exceptionality and originality of writing is almost bordering on one of the criteria.”

Theme 2: Belief

Before the focus group discussion started, a presentation was given on how XIP works, what research had been carried out until that time, and its results. Additionally, participants

were told that the aim of this research is to support their essay assessment practices, not to automate marking. Participants were asked to write down their initial thoughts on paper regarding the whole idea of using an automated text analysis tool to support essay assessment practices. Their initial reactions, however, demonstrated their belief about what automatic technologies are capable of, which was still around the idea of automating the assessment and marking. Here are couple of examples of the initial thoughts:

> PARTICIPANT 2: "Can't help being suspicious that in the end a better version of XIP will be used to automate assessment!"
>
> PARTICIPANT 3: "Worry of its use for summative assessment…"
>
> PARTICIPANT 6: "I am always reluctant to rely on a machine to make a judgement about a human activity like writing."

They were then asked to write their final thoughts after the focus group session finished. When initial and final thoughts were compared, there was an obvious increase in eagerness to use the tool, none of them rejected the idea of using such automated support, even though potential problems had been identified during the discussion. Their initial thoughts were concerned with to what extent they could trust a machine; their final thoughts focused on motivations for its use. The participants began with preconceptions about automated technologies, but as they understood more they became more interested. Here are some examples which illustrate their final thoughts:

> PARTICIPANT 2: "If we could find a way to make it useful for the teacher to work with the student, that could be great."
>
> PARTICIPANT 3: "Great to help student writers to develop clarity of argument in their academic writing. Would use as a discussion tool with students and would see as valuable tool for students learning to edit their own work provided limitations were made clear."
>
> PARTICIPANT 6: "A good idea and a tool. I would use it if it was adapted to analyse at a more complex level."

A comparison of the initial and final thoughts demonstrated a shift in participants' opinion regarding what they believed an automated text analysis is capable of. To begin with, participants criticised such automated technologies since they believed that such tools were developed for commercial return, and not to support assessment practices or markers. Additionally, they stated their belief that automated marking could never be as good as human marking; whatever advancements were made, it will still be a machine.

> PARTICIPANT 4: "this is obviously – I mean whatever we criticise it is obviously something that [1:08:44.4?inaudible] commercial world-"
>
> PARTICIPANT 2: "And at the end the truth is that the people developing these tools, they want commercial return for them, so the more delicate the more niche and therefore the less lucrative unfortunately."
>
> PARTICIPANT 6: "I know probably sounds slightly apocryphal, but years ago I read the comment that artificial intelligence is no match for natural stupidity. And I'm not trying to be flippant but the thought behind it is that a machine on a programme, which functions in binary, which is basically positive or negative, no matter how sophisticated it gets; no matter how far or nor far you go back to find a human being programming it, it is still – it's still a device that is logical."

Although participants initially approached the research assuming that eventually marking would become automated, in the end they considered how they would use it to support their assessment practices. Changes in their opinion occurred at different points in the focus group. For instance, at the end of the presentation, once the quantitative research results were discussed, participants' opinion towards the overall research became positive.

> PARTICIPANT 2: "Oh wow. I'm really impressed, really!"
>
> PARTICIPANT 4: "That's great, ambitious, really interesting."
>
> PARTICIPANT 1: "I feel like it's really developed, really developed since last time."

PARTICIPANT 5: “You’ve done so much research and it’s so ‘together’.”

PARTICIPANT 3: “And it’s the dialogue that’s so valuable, isn’t it, because obviously there are lots of people working from the computing side who have got ways of approaching but perhaps don’t have the categories that the linguists would claim.”

After completing the highlighting activity, the XIP analysis results were given to the participants. Each received the printed XIP analysis version of the piece that he or she had just highlighted. The general reaction was very positive. Participants expressed their surprise and showed great enthusiasm when they looked at the page in the first instance. Participants were surprised by the extent to which XIP analysis resembled theirs when they compared it with their own judgement and this triggered the shift in their opinion.

PARTICIPANT 4: “Oh right, how interesting, oh! ((*Laughs*)) Actually, they are not that different.”

PARTICIPANT 3: “Wow.”

PARTICIPANT 2: “It is a really strong start isn’t it?”

PARTICIPANT 6: “It’s not different at all.”

Once participants had been introduced to what the tool is capable of, quantitative research results were discussed. When they compared and realised the similarity of the tool’s results with their own judgement, their initial opinion changed. At the end of the whole session, participants left the room with following closing remarks:

PARTICIPANT 5: “It’s interesting. I always thought it was going to feel ridiculous today; but it’s been really beneficial.”

PARTICIPANT 2: “I realise now it is beneficial, yes-”

PARTICIPANT 3: “…great - really made me think about practice of teaching and marking academic writing! Thank you - but we are such a critical lot aren't we? You know we would criticise anything - even perfection - but so impressed by the

way you organised us in the tight timescale you had and what an amazing amount of work you have done -- I really enjoyed - well worth travelling down from (…)"

The highlighting activity, explained in section 7.3, revealed another belief about how the participants viewed student writing. Although the participants were told that they should highlight a full sentence for comparison later, they found this hard and one of the participants even resisted this idea and simply highlighted the key words or phrases that s/he liked. The reason for this was because participants were not used to highlighting the sentences fully, as they were accustomed to looking at the text on a phrase level, and they did not found everything salient in one sentence.

PARTICIPANT 2: "Could I ask, do you want whole sentences because I mean I could only come at it at the phrasing level -"

PARTICIPANT 3: "You can underline the sentence in which the thing that you've identified is. That is what I have done."

PARTICIPANT 5: "I want to go that bit and that bit rather than the whole thing."

Although the participants looked at the text at a phrasal level when highlighting, they also looked at the example text as a whole. When the participants assigned XIP categories to each sentence, they were not talking about one particular sentence at a time. Although the activity involved labelling individual sentences, they tended to look at the bigger picture and interpret the deeper meaning, based on the previous or next sentence, even paragraph. The following dialogue is given as an example:

PARTICIPANT 4: "I was thinking- you can't look at that paragraph without the next paragraph"

PARTICIPANT 2: "I mean when the tool was developed, it was developed by people who haven't even thought that the sentence might not be the unit –the unit of analysis that you should have started with?"

PARTICIPANT 4: "Or the word."

PARTICIPANT 2: “Or the word – because those are the kind of – its textual structure-”

PARTICIPANT 4: “-a clause, a clause or phrase is one element you’ve missed out and the other one is text.”

PARTICIPANT 5: “It’s all grist to the mill of thesis, yeah.”

This dialogue demonstrates that participants look at student writing from all aspects, its structure, words, and phrases, and check out how these elements form together and how well these elements tie together to create a whole essay. Therefore, they do not simply assess a paragraph in itself but look at the next paragraph too, the bigger picture.

Theme 3: Power and politics

Prior to the focus group session, during the presentation on how XIP works, and considering what research had been carried out up until that time (chapter 5), the participants were told about the outliers identified in the results of that research, (as given in section 5.5.5.1). Up until this point, the participants listened to the presentation carefully with no reaction, even to the parts where the tool’s successes and strengths were identified; but when participants discovered problems with the XIP tool and were introduced to the outliers, they were happier and pleased.

DB: “… and similarly there were low graded essays with a high number of salient sentences.”

((*A-ha*, *Hum*))

DB: “Like in this sentence for instance the student was just giving an example from a Peter Pan story saying that ‘Wendy is not seen to challenge this role’; but because of this challenge word it is – XIP thought it’s a contrast – so it didn’t work.”

((General agreement))

PARTICIPANTS 1&3: ((*cynical smiling, looking at each other, nodding*)

PARTICIPANT 4: “I see!”

PARTICIPANT 2: ((*eyebrows up, cynical smiling*))

As mentioned before, after completing the highlighting activity, the XIP analysis results were given to the participants. Each received the printed XIP analysis version of the piece that he or she had just highlighted. Participants were surprised by the extent to which the XIP analysis resembled theirs and this shifted their opinion in a positive way. Their enthusiasm, however, was followed by a greater desire to discover where the tool might have failed. The participants were more enthusiastic about finding issues with the tool and identifying its weaknesses.

PARTICIPANT 1: "Well the first thing I'm surprised about is that the word 'while' hasn't triggered contrast. ((General agreement)) Why is that? Because 'while' is a classic way of signalling – I mean that's just something missing in the tool.

PARTICIPANT 5: "This one is a really good one to say it really doesn't work. It isn't a contrast, it just isn't."

During the focus group session, the participants questioned how far automated text analysis could go and what would happen to teaching proficiency as a result of that. They emphasised how markers fear losing their jobs and being replaced with automated technologies in the future.

PARTICIPANT 5: "Can I just ask, I have a question? In looking at what's out there and what people are doing already, have you looked at all at what the TOEFL examination boards are using? They were trying to introduce something very similar at least five years ago so I don't know how far down the line they've got."

PARTICIPANT 1: "But that's why, hence all of those complex political issues around recruiting for your study because some of the ALs out there you were asking to take part would be thinking, yeah, that sort of stuff is going to replace me in a few years, yeah, yeah."

PARTICIPANT 3: “That’s a fear that they will be completely replaced. I know that they were asking the TOEFL researchers to look at that very question, how far the markers could be replaced?”

(from Initial Reactions) PARTICIPANT 2: “Can’t help being suspicious that in the end a better version of XIP will be used to automate assessment.”

When the participants discussed how they would use the XIP tool, potential problems and limitations were also discussed. The participants talked about possible conversations they could have with their students and what future scenarios could look like. The participants also emphasised that they would like to take control over such tools. Instead of training tools to mark essays and to bring them up to the quality that equals theirs, they preferred to ‘harness’ them, so as to feel superior over such technologies and to control them to obtain benefit.

PARTICIPANT 1: “I can see students saying, oh it says this, so why haven’t you given me a good mark? I can see that. Because this says I’ve used the arguments – so limits have to go with it.”

PARTICIPANT 6: “And, and anything computer related or which has got any numerical element to it has a tendency to suggest that there’s a perfectible – something that eventually will approximate to an accurate mark and I think if students think that something – that there’s a correct mark, they will always go on finding the human beings marking their assignments wanting. And there will be just too many unnecessary conversations about why they haven’t got the mark they -”

PARTICIPANT 2: “-definitely with the limitations of it that this doesn’t mean your tutor is going to mark these things, this is to help you edit your work, that sort of thing.”

PARTICIPANT 5: “And also there’s the risk that to some extent what (. . .) was talking about that you teach to the marking tool, which is a complete anathema really isn’t it? It’s not the way we want to go.”

PARTICIPANT 4: “But if there was a way of harnessing it so it enabled you to do your job as well or better and quicker, that would be wonderful.”

PARTICIPANT 2: “… where there’s an assessment relationship it’s got to be made clear to a student that this is not – the machine knows the right answer and the human beings aren’t matching up to the machine.”

PARTICIPANT 1: “… this doesn’t mean your tutor is going to mark these things, this is to help you edit your work, that sort of thing.”

PARTICIPANT 6: “I am always reluctant to rely on a machine to make a judgement about a human activity like writing. I much prefer the machine to assist me in arriving at a judgement.”

Theme 4: Problems

During the focus group discussions, participants talked about some problems that markers and their students experience which potentially automated support could solve. One of the problems they raised was about the reliability aspect of human marking. They discussed and admitted that there is a mismatch around human marking. Consider the following script which illustrates their concerns about subjectivity in their marking:

PARTICIPANT 2: “There’s a widespread tendency to try – to – for us all collectively pretend that assessment in higher education is capable of becoming, of being objective, or that we can approximate to objectivity and the honest truth is that we cannot, we cannot.”

PARTICIPANT 1: “Yes, yes, we know that-”

PARTICIPANT 4: “Yes, I was going to say it isn’t the sort of like the human marking is flawless-”

PARTICIPANT 2: “So it just has to be treated – unfortunately obviously universities are constantly -, they are encouraging the notion of objectivity because students want it, because they think it’s fair. So we don’t tell students the half of the truth about how- just how subjective the process is. We do our best to make it fair but you cannot eliminate human – the human variations.”

PARTICIPANT 3: “Really interesting; how markers might be prejudiced.”

PARTICIPANT 2: “Yes, but none of that, that’s not understood at policy level at all – so you have this idea that if you mark someone’s work honestly that that somehow means you are being objective, you know.”

Participants admitted the fact that human marking is not flawless, and that there might be variations and subjectivity issues. The validity issue had also been observed while participants were discussing their decisions about which rhetorical moves and discourse elements constitute a positive mark for student writing. There were some different approaches, not only in the way they assess argumentation, but also in the way they decide when to credit and award a mark, which sentence deserves a credit for, say, ‘alternative explanations’. To illustrate this point, see the following conversation:

PARTICIPANT 5: “I wondered about the question at the end of the second paragraph, ‘this prompts the question’ because isn’t that the beginning of the argument? I underlined that.”

PARTICIPANT 4: “I don’t think rhetorical questions generally are the most sophisticated way of setting up an argument. But if you see a student doing that I think your immediate response is they are aware of alternative explanations … so the student is probably going to pass on this in the essay.”

PARTICIPANT 6: “But I wouldn’t give much credit for it if it didn’t then.”

PARTICIPANT 5: “But there is at least some awareness.”

The participants also pointed that most of the markers are not linguists, therefore they may not necessarily notice that their students are making actually an analytical point and therefore they might under-mark essays. Additionally, markers only have a limited time to mark an essay, hence they may not be able to study these linguistic features closely. Even though rhetorical features, such as the use of 'on the other hand', are sometimes provided in marking rubrics to give credit, markers might not credit subtler rhetorical formulae as they need to work on the margins of the rubric. Therefore, potentially, it might be valuable to use an automated text analysis technology that identifies such meta-discourse features, the linguistic features that show to the marker analytical points within student writing.

> PARTICIPANT 3: "And I, I think that it is arguable that even though some – that people are sometimes under-marked; I've been doing some monitoring in the last couple of weeks and I've been monitoring things that I think have been slightly under-marked partly because I think the – the ALs aren't necessarily noticing when a student is making actually an analytical point; possibly, you could speculate, because the student isn't using this meta-language that makes it clear that they are being analytical."
>
> PARTICIPANT 2: "I can still see why it might be saying something valuable in terms – because – in a sense because markers aren't linguists. Because markers aren't able to study things closely and they don't have the time."
>
> PARTICIPANT 1: "The time as well, especially with the marking."
>
> PARTICIPANT 2: "And if I could just add to that, in modern language marking, these rhetorical formulae are actually in the marking criteria; so the marker has no option but to give credit for the 'on the other hand', 'on this hand or the other hand' kind of argumentation and cannot in a way give the same credit for something more subtle. So when you build it into the criteria you create a different problem."

When the participants were making informed judgements about possible XIP categories for each sentence, there was an obvious difference in the way they interpret a sentence (see table 7.7). There was no agreement in the naming of what would be the sentence type; all six participants were throwing out different opinions. Although it is true that their thinking is complex since they look at the text as a whole (as given in theme 2: belief), their thinking is not always the same. Their decisions differ from each other in the same way as they do from XIP. The participants thought that this difference was because human beings are much less sure about what it is they are looking at. Therefore, similar discussion could be used pedagogically with their students to make them think about their writing.

> PARTICIPANT 4: "Well but I think what we've come up with here is in itself a finding isn't it? In that human beings are much less sure about what it is they are looking at than the machine is and that – I think we have to assume that the machine is a bit of a blunt instrument in that case."
>
> PARTICIPANT 2: "But I think it could be used pedagogically couldn't it because this is such good stuff of debate." ((*General agreement*))
>
> PARTICIPANT 1: "Yes, I mean if it were possible to use this to get students to think about their writing –"
>
> PARTICIPANT 5: "I am thinking I could use it in a, yeah, in an exam tutorial certainly."

The participants also mentioned that such an automated support could be used to improve the quality of their feedback, to make sure they gave a clearer explanation about why they have given a specific feedback to their students.

> PARTICIPANT 6: "I'm a brand new tutor on a third level English Grammar course and it's – and one of the things that really would help students sometimes is to know why that has been given as the answer. I sometimes spend hours a day explaining well it could be this and it could be that but it was felt that was the more

likely, but that is an equally valid alternative. However, and I'm not being at all critical, it's just that sometimes it's that -"

PARTICIPANT 4: "Absolutely, I agree."

PARTICIPANT 6: "But at – that kind of level we are expecting them to think critically and ask the questions."

PARTICIPANT 4: "A self-help diagnosis kind of thing."

PARTICIPANT 6: "A failsafe mechanism that ensures I haven't overlooked anything."

PARTICIPANT 2: "I think it could also be used in the context of tutor feedback. So if you were trying to – I mean one of the labour intensive things about marking is connecting your general comments with specific in situ examples. You know, because it is actually quite time consuming to say, you need to do X, Y and Z – for example, and then to set out an example. And some tutors use like cross-references and things.

If you had something automated that you could say, I've highlighted – or look at your assignment as it's been highlighted by the tool, you will see that at the moment there isn't much argumentation in there, there's a lot of description or something like that. But obviously you'd have to see what the output looked like before you sent it to the student in case it wasn't any use."

The general agreement was that there is potential to use the tool as part of a formative assessment, but definitely not for summative. The participants also agreed that the tool can be used not only as a self-help with students in the context of participant feedback, generating discussions and teaching; but it can also be used by students themselves as a way of reflecting on, critiquing and editing their work; which has some potential risks.

PARTICIPANT 5: "– quite a lot of my level two students just do not edit, they do not edit at all. I mean they might spell check, okay, but they don't edit and I would

see this as a kind of starting tool. Where do you start to edit your work? Well here, have a go at this and see what it brings out if you don't use argument – that sort of thing as a sort of self – put it through the tool."

PARTICIPANT 2: "I think that point is really important (…) in that it's potentially a tool not just for tutors but for students."

PARTICIPANT 5: "Danger that students might start to use the meta-discourse markers uncritically. This could lead to bland writing without understanding that the meta-discourse is a visible sign of thought and argumentation."

PARTICIPANT 4: "So looking at these words, it's not – it doesn't mean that you are going to get higher marks necessarily if you use them but we are just looking at it as a tool to help you to edit and criticise your own work; as long as it was really transparent what it was doing and the limits of it, then I think it would be very useful."

PARTICIPANT 6: "You would sort of use it to generate discussion. You would say, okay, take a piece of – not people within the room, but you would get it to analyse and then use it -. I mean it will work, you know, as long as you have the questions, it would definitely work to generate discussion."

Theme 5: Changes

From the focus group discussions, it emerged that participants would like to see some changes made in XIP before using it to support markers in essay assessment practices, some of which are deep changes and others are superficial.

Deep Changes

The deep changes that participants would like to see made were about the tool's theoretical underpinning, the tool's sentence-level analysis, and the tool's categories.

- **Theoretical Underpinning**

During the presentation about XIP, prior to the focus group session, the participants wanted to know more about the linguistic theories that were used when creating the XIP tool.

> PARTICIPANT 1: "Can I just, I don't want to interrupt long but can you just remind us what the theoretical, was there a theoretical background?"
>
> DB: "Yes, yes, I was coming to that."
>
> PARTICIPANT 1: "Perfect, 'cause you know it is important to us."
>
> DB: "Without any explicit linguistic theory behind it…"
>
> PARTICIPANT 1: Aha! ((*eyebrows up*))
>
> PARTICIPANT 4: "You are not going to convince linguists until there are some really principled and solid-sounding empirical basis for that..."

The participants were hoping for a theoretical basis for XIP's approach. As set out in the literature review, there is no explicitly stated linguistic theory behind XIP at the moment. However, although the underpinning theory was not linked to an existing linguistic theory, some theoretical works have links to XIP as given in chapter 3.8.4.

- **Sentence-level analysis**

While the participants were comparing their highlights with the XIP analysis, there was general agreement that the tool was missing some obvious CONTRAST statements.

> PARTICIPANT 2: "The second 'however' is highlighted further down the page but not the first one."
>
> PARTICIPANT 5: "((reading the sentence)) - 'However a convincing case for a child's ability to ignore all but most transparent has been intended' – that is definitely contrast, how come the tool doesn't pick that up?"

PARTICIPANT 6: “Well the first ‘however’, which is two-thirds of the way down the second paragraph is starting a new sentence and before you were giving an example of how an argument can be staged over. We teach our students the value of sometimes breaking down their arguments into separate; … yet it sounds like the tool doesn’t want you to do that-”

The participants emphasised that due to the problems originating from the sentence-level analysis behaviour of XIP, the assigned XIP categories might not be correct. To illustrate:

PARTICIPANT 5: “Where’s the contrast? I think this is a fundamental issue with it because it’s picking up a word like ‘argues’ and it’s assuming that within that sentence is, therefore, going to be some kind of contrast.”

PARTICIPANT 2: “Exactly – and this is only half an argument anyway because prior to that is ‘Peter Hunt says this, Reynolds says this’ and in fact both of them therefore say that. So it’s actually, there isn’t contrast.”

- **Categories**

Although the XIP categories, their definitions and example sentences for each category were introduced and explained in the presentation session prior to the focus group, the participants struggled to make sense of why XIP assigned particular categories to some sentences.

PARTICIPANT 4: “You are not going to convince linguists until there are some really principled and solid-sounding empirical basis for that. It is really difficult to see how, what is that the machine is picking up there.”

Although the reason was mainly because of the sentence-level analysis problem, the participants also raised the issue that XIP needs to unpack the underlying meaning of each category. Their points suggested changing, adding or revising some of the categories. For instance, participants suggested the category CONTRAST should be unpacked, since what had been assigned by XIP was not

actually suitable for this category. The student was not contrasting the idea but endorsing what had already been said by another researcher.

> ((The section participants talk about is written in two separate consecutive sentences:
> 1. Peter Hunt highlights how children are…
> 2. In addition, Kimberley Reynolds argues in favour of …, claiming …))
>
> PARTICIPANT 1: "It's the opposite of contrast, it's endorsement. ((talking about the second sentence))"
>
> PARTICIPANT 5: "This one is a really good one to say it really doesn't work. It isn't a contrast, it just isn't."

Another issue raised during the discussions was that XIP works lexically, whilst they felt that it needs to be based on semantics. Particularly during the activity where the participants labelled each sentence with an XIP category, there were lingering concerns about the purely lexical identification of rhetorical devices as opposed to semantic. They all concluded that the level of analysis needs to move beyond the lexical to the phrasal/clausal/textual.

> ((The sentence participants talk about is:
> At the heart of this debate lies the challenge of defining the purpose of children's …))
>
> PARTICIPANT 5: "Well because it's lexical- because it's gone for 'debate', 'challenge', 'purpose' and 'focus': well you could say that was EMPHASIS or SUMMARY couldn't you? Debate, challenge, purpose and focus; so debate and challenge are the only ones that are contrastive."
>
> PARTICIPANT 2: "I think it must lie at the root of how they've developed the tool; how they came up with-"
>
> PARTICIPANT 5: "This is all lexis; it is not semantic, and it is not grammatical; it is lexical."

PARTICIPANT 2: “I’d say move away from lexis to higher levels of analysis.”

((General agreement))

PARTICIPANT 1: “I would say that too – move it away from lexis, I would say that, yeah.”

Superficial Changes

The superficial changes that participants would like to see made were about building the tool into a dashboard, supplying it with a manual, and providing them with the ability to customise it.

- **Built into dashboard**

  During discussions on ways in which the XIP could be used, the notion emerged of building it into a dashboard. The participants suggested two different dashboards: one for markers and educators, who could input student writing and see what categories the tool assigns, which could then be used to base their feedback on; and another for students who could input their work for revision purposes.

  PARTICIPANT 5: “So this is – so I have a student and I put their work into the tool and the tool – gives me a number of contrasts or emphasis – So I suppose what I could use that for is to use that to base my feedback –”

  PARTICIPANT 1: “You know, on the website or, you know, you could have it customised for your own course’s website.”

  PARTICIPANT 3: “I think this would be useful there and I’ve got, you know, make it student friendly so they can put things in.”

  PARTICIPANT 1: “I think I would use it as a self-help with students; I would like to see something like that on a student’s dashboard where you’ve got, you can put your stuff in - in for argumentative language and

you can just – as a means to help students to edit with complete understanding that -"

- **Provision of a manual**

The participants noted that XIP should explain why it selects certain features. It should be made easier to use if there are clear explanations of what its results do/do not mean. The meaning should be transparent which means clearly setting out what it does, and what its limits are so that it can be useful.

PARTICIPANT 2: "- explanations of what the data does and doesn't mean, alongside the display of the data."

PARTICIPANT 3: "– make transparent what you are getting out of it."

PARTICIPANT 4: "Explain the logistics of the data about what, what the implications of -"

PARTICIPANT 2: "An ability to sort of contextualise what your – yeah."

- **Ability to customise**

The participants emphasised that it should be possible for users to be able to customise the tool, so it can be adapted for use at different levels, and for different disciplines and modules.

PARTICIPANT 6: "Well it certainly needs to be adaptable to the data that you are using. It is not going to help anybody – I think if it attempts to be too generic it will fall between all the different stools."

PARTICIPANT 2: "-ability to adapt or select for my discipline and module; so that I could use it at the beginning to make some kind of effort to tailor what the results would do to what I'm putting into it in the first place. So discipline or level, or empirical work or discussion or review or – if there were some way of, in an ideal world if you could sort of navigate to a particular set of categories that were more likely to apply to your –"

DB: “So you mean before let’s say uploading your student essay you would like to choose the level of the student, the field of this essay and even if it’s an essay or a technical report or whatever it is, so you need to choose this?”

PARTICIPANT 1: “Yes, so to customise it.” ((General agreement))

PARTICIPANT 2: “The level is really important actually; we’ve not discussed that before because how you would mark a level one compared to how you would then approach a thesis.”

7.6.1 Discussion

The focus group study brings a new dimension to this research as it offers an understanding of how to use the XIP tool for supporting the process of assessment. It is important that participants trust a machine that is analysing human writing, and therefore it was important to hear what queries or even doubts they had about how such a tool works, as well as how similar its output is to their judgement of quality, and how it can be improved. The one-to-one interview markers, as discussed in chapter 6, provided answers to how educators define the attributes of good quality student writing, and to what degree XIP can identify the presence of these attributes. Before the focus group study, there was no evidence about what participants would like/dislike about XIP, how they would react to the concept of integrating a machine as part of their assessment, and whether the XIP output is congruent with the participants’ judgement of quality.

The focus group study helped to move the results of previous studies further, not only by triangulating its results, but also by offering multiple perspectives on exploring the possibilities of integrating the XIP tool into educational contexts to analyse undergraduate writing and to support essay assessment practices.

Five overarching themes were found in the data analysis for this study:

Theme 1: Quality

The participants pointed to aspects of subtlety, form, beauty and flow as signs of good quality student writing. This finding is not congruent with the academic writing literature, since exceptionality in student writing has not been given as sign of quality in prior research. It has to be noted, however, that the prior research on teaching academic writing, conventions of academic writing and what makes good quality student writing was only reviewed through the literature from English-speaking countries. Other countries, such as China, make the use of allusion, analogy, and proverbs as one of the important criteria for grading writing since their usage is thought to contribute to the beauty of the language (Xing, Wang, & Spencer, 2008). In Chinese, the beauty of writing is believed to lie in delicacy and subtlety (Shen & Yao, 1999). However, the literature on academic writing from English-speaking countries does not consider beauty and subtlety to be signs of quality. The finding is congruent, on the other hand, with the McNamara *et al*. (2009) study regarding the definition of writing quality based on human judgments, chapter 3.6. They found that higher scored essays were more likely to contain linguistic features associated with text difficulty and sophisticated language, thus concluded that more skilled writers use more sophisticated language.

Additionally, although the participants pointed to beauty, flow and subtlety as signs of quality, this finding is not congruent with the one-to-one interviews, as given in chapter 6 (section 6.6). None of the tutors, interviewed individually, mentioned beauty as sign of good quality student writing. This could be explained by the fact that in the one-to-one interviews tutors used specific marking rubrics, which did not include beauty and subtlety as signs of quality. This is in line with what the focus group participants also commented, that these are not necessarily taken into account in marking rubrics as signs of quality.

The focus group participants stated that some student writings show different literary styles which are not common but exceptionally good. For instance, they mentioned how

exceptionally written the example text was. The student was using impressive and original phrases and words that needed to be credited such as: ‘the question of whether’, ‘a breeding ground for’, ‘by extrapolation’, ‘a convincing case has been contended’. In fact, XIP highlighted all the sentences containing these phrases, which showed in this context and example that the XIP analysis of this student essay was congruent with the educators’ judgement of quality.

Theme 2: Belief

The participants came to the focus group session with preconceptions about automated technologies. Their belief was that such technologies were developed for commercial return, and that they can never be as good as human markers; the aim of using automatic technologies is to automate the marking, not to support assessment processes. However, a comparison of their initial and final thoughts demonstrated a change in the participants’ opinion regarding what they believe an automated text analysis is capable of. Their initial thoughts were concerned with trusting a machine; their final thoughts focused on motivating its use.

Changes in their opinion occurred at different points. For instance, after the presentation session on what XIP does, learning about what had been already found in prior research in the quantitative studies (chapter 5) changed the participants’ opinions positively towards the overall research. As they understood more, they became more interested. Similarly, after completing the highlighting activity, when participants examined the XIP analysis and compared them with their own highlights, they were impressed. This shows that they did not expect the analysis of the tool to resemble their own decisions. When the participants found out the XIP highlights were congruent with their highlights, their attitude was very positive which was a shift from their initial opinion.

The theme belief also revealed how participants approach assessing student writing. For instance, the participants found highlighting at full sentence level hard and preferred to

highlight phrases. The reason for that was because the participants were not used to highlighting the sentences fully, they were accustomed to looking at the text at a phrase level. Additionally, the participants looked at student writing from all aspects: its structure, words, phrases, what these elements form together, and how well these elements tie together to create a whole essay. Therefore, they did not simply assess a paragraph in itself but looked at the next paragraph too, the bigger picture. This finding is in line with the prior study, as given in chapter 6. When the natural behaviour of tutors was observed during the one-to-one interviews, tutors tended to develop their own styles, i.e. reading the whole text in one go, highlighting mostly at phrasal level, and checking whether students had achieved what they claimed to do so before crediting it.

Theme 3: Power and politics

The participants were happier to discuss XIP's weaker points than they were identifying its strengths, which implies that they did not want its quality to equal theirs. The theme power and politics emerged since the participants wanted to feel superior to automated technologies, 'harness' and control them to obtain benefit.

As stated in page 233, emerging from the focus group discussions, the underlying issues of power and politics were due to the participants' fear of:

- what might happen to the future of the teaching profession and them losing their jobs as a result of that
- and being judged by their students who could potentially compare human and machine results.

This finding triangulates with the literature (Attali & Burstein, 2006; Elliot & Williamson, 2013; Ericsson & Haswell, 2006; Shermis & Burstein, 2013) indicating the suspicion and tension of writing teachers and researchers towards the use of automated text analysis. It also triangulates and tallies with the one-to-one interviews, where tutors felt uncomfortable

discussing how technology might help to overcome the problems they experience with assessment, and stated the worry that they might be replaced by technology.

Theme 4: Problems

The participants talked about some problems that markers and their students experience which automated support could potentially solve. These were identified as:

- the subjectivity of human assessment and marking
- the limited time that markers have to assess an essay
- the possibility that markers do not necessarily notice that their students are actually making an analytical point, since most of the markers are not linguists
- markers need to improve the quality of their feedback and make sure they give a clear explanation of why they have given a specific feedback to students
- markers need to generate discussion with students who are required to reflect on, critique and edit their work.

Some of the problems identified tally well with the one-to-one interviews and are congruent with the literature. The problems of subjectivity in human marking, time limitations, and the need to provide better feedback and examples to ensure students understand their reasoning, became evident in the one-to-one interviews with tutors. Earlier research (Lea and Street (1998)) indicating that academic tutors experience difficulty with providing effective examples and feedback is also supported by the focus group findings.

These problems, such as the labour intensive, time-consuming essay assessment problem, could potentially be solved through automated support. Considering it took around 15 minutes for the focus group participants to highlight 13 sentences, it could be time-efficient to use the automated support as the XIP analysis for example took less than a minute for the same piece. Additionally, automated support potentially could help to overcome the subjectivity of human marking. Participants were honest about how

subjective their marking can be and that there was a mismatch between the way they interpret and mark things. This is a very critical point showing the possible inconsistencies between human markers. In line with earlier research (section 6.4), human markers can disagree with each other and therefore they do not necessarily come to the same conclusion as their peers; which is a reliability concern. The automated output on the other hand is always the same, stays the same and is not subjective. The XIP tool could therefore potentially be useful to help educators overcome this problem by using it as a self-reflection tool. Once brought up to the correct level based on the suggested changes, there is a possibility that it can offer educators a self-diagnosis to reflect on their assessment and marking.

Theme 5: Changes

It emerged that participants would like to see some changes made to XIP before using it to support markers in essay assessment practices, which would help the XIP to become more congruent with educators' views of quality. Some of these are **deep changes** (theoretical underpinning, sentence-level analysis and categories) and others are **superficial** (built-in dashboard, supplying a manual and being able to customise its use).

The first deep change that participants requested was in respect of the theoretical underpinning. Although XIP is not underpinned by a specific linguistic theory, as given in chapter 3.8.4, it has been developed based on empirical evidence and linguistic concepts. Even though the participants questioned the existence of an underpinning linguistic theory, this did not ultimately hold them back from feeling positive about the idea of using to the tool and from discussing how they would use such a tool at the end of the focus group session.

Secondly, the participants raised their concerns around XIP's sentence-based analysis. They emphasised that since XIP does not assign labels considering previous or following

sentences, the XIP categories assigned might not be correct. This is congruent with the outlier analysis as given in chapter 5. As shown there, the XIP tool analyses a document sentence by sentence and so cannot identify a classification such as CONTRAST across several sentences. Consider the following example that shows how the expression of CONTRAST is distributed across several sentences (underlined):

> As Hunt states 'sameness and difference is the essence of children's books; they have recurrent ideas' (2009a, p. 71). *He goes on to cite …* [*Here comes a list of examples*.] But is this the only tradition the book breaks? Based upon the themes detailed above this essay will look at what similarities and differences A Monster Calls has to children's literature from the last 250 years, focusing particularly on Tom's Midnight Garden.

As XIP works at a sentence level, an argument made up of three consecutive sentences will not be picked up. This is one way in which the accuracy of XIP is shown to be limited. Thirdly, since the participants struggled to make sense of why XIP had assigned particular categories to some sentences, they suggested unpacking the underlying meaning of each category. Their points suggested changing, adding or revising some of the categories. This is congruent with the findings of chapter 6 (one-to-one interviews with markers) that XIP needs additional categories and it needs to unpack the underlying meaning of each category (section 6.8). As the student writing in the example given shows, although the student was using the word 'argues', s/he was approving the idea argued by someone else instead of implying any CONTRAST. Therefore, the focus group participants suggested a new category name, ENDORSEMENT, for this particular sentence type. XIP needs to unpack the underlying meaning of each category, and additional categories need to be adapted for educators' language.

Three superficial changes were suggested regarding the ways in which the XIP tool could be improved. These changes are possible although they are not straightforward, and requires empirical research and time. The XIP's actual output does not provide any explanation regarding the meaning of the data or the tool's limitations. However, this can

be presented in an environment to its users, in which XIP is embedded and where its pedagogical grounds are made clear.

7.7 Conclusion

The focus group study showed that although the XIP tool needs to be improved and its limitations and risks should be acknowledged, XIP impressed the focus group participants, who were not initially inclined to use an automated technology to analyse student writing. It showed them that such an automated rhetorical parser could be useful, although their initial thoughts were the complete opposite. This study moves the research forward regarding what should be done to bring automated meta-discourse analysis usefully into the lives of the participants without them needing to be concerned that it would eventually take over their roles. This study prompts consideration of how human markers and machines can work well together and mutually complement each other for their own sake and for their students. Advancing automated support for assessment is key when the strengths of both sides can be brought together: the speed and reliability of the machines, and the vast capabilities and the knowledge of the human markers.

In conclusion, this study suggests that the congruency exists, as demonstrated by the participants who left the session feeling impressed and positive regarding the idea of using an automated technology to support their assessment. Their initial suspicions had changed, and their final thoughts were about how the tool could be used effectively. XIP offers possibilities, but it is not the right tool at the moment for those who are not initially inclined to use a machine to analyse student writing. However, this study points to what the right tool could offer. There is a human element – beauty, subtlety, truth, meaning – which today's technology cannot capture, but there are also tools that people need to help with their writing and to help with their marking, that can provide consistent and reliable results.

Although potential risks and limitations of using XIP were talked through during the discussions, the final thoughts of the participants were positive towards using XIP and there was general agreement about the idea that once the XIP tool has been improved in the suggested ways, it could support both educators and students. Although the current level of congruency between XIP and the educators is low, this study proposed ways of making it congruent through potential changes to XIP; some of those proposals were deep and some were superficial, as explained above. The next chapter describes how the XIP tool could be brought up to this level and what sort of environment could be created for educators and students to support their assessment.

8

DELIVERING THE XIP OUTPUT

8.1 Introduction

This chapter addresses the fourth subsidiary research question: "In what ways should the output of the XIP analysis of student essays be delivered to educators so that they make use of this output to give feedback on student essays?"

The initial attempt to answer this research question was made by means of a pilot study carried out during the first year of this PhD study, which is explained in next section, 8.2. The primary aim at that time was broadly to explore ways in which the XIP output could be visualised in a more user-friendly way. Following the pilot study, later in this PhD study, the focus turned to possible ways of visualising the XIP analysis of student essays for supporting educators' assessment practices. The main study, which is explained in section 8.3, addresses the fourth subsidiary research question with this focus.

8.2 Pilot Study

8.2.1 Motivation

Based on the reasons given in section 3.9.1, the XIP has been used as an exemplar computational language analysis tool to answer the research questions of this thesis. Once this decision was made, the initial attempt was to think of ways in which the XIP output can be shown in a user-friendlier way, because, as is typical of language technologies, the XIP generates a semantically tagged XML file as an output suitable for subsequent

machine analysis. To illustrate, as shown in the following figure 8.1, the XIP extracted three rhetorically salient sentences from the document that was analysed (IMPSENT-important sentence). These important sentences were labelled as CONTRAST, CONTRAST and SUMMARY because of the meta-discourse (_MD+) markers identified by the XIP (words followed by _MD+ tag). For detailed explanations of the XIP categories, please see section 3.8.3. Additionally, for each rhetorically salient sentence, key terms mentioned within the extracted sentence, shown as CONCEPT (i.e. 'large industrial scale' is a key term, CONCEPT, for the first important sentence whose rhetorical move assigned as CONTRAST through the meta-discourse cues of *problem* and *method*).

IMPSENT_CONTRAST (The problem_MD:+ with this method_MD:+ is a lot of electricity is required especially on a large industrial scale, which would be required on Mars.)
CONCENT_IMPSENT (large industrial scale)

IMPSENT_CONTRAST (Another problem_MD:+ with this method_MD:+ is that it is not efficient and produces a large amount of wasted heat.)
CONCENT_IMPSENT (large amount of wasted heat)

IMPSENT_SUMMARY (In_MD:+ conclusion_MD:+ I would suggest_MD: + that such a method_MD:+ would be_MD:+ a useful method_MD: + of searching for water supplies.)
CONCENT_IMPSENT (water supplies)

Figure 8.1 Raw XIP output

While such plain textual output could be fine for and well-suited to researchers, computer scientists, or linguists to analyse it manually, or with other tools, this is definitely not a form which could be usefully or attractively presented back to learners and educators seeking to assess their writing. Therefore, the need to visualise XIP's promising results in more user-friendly ways was recognised.

Prior to this PhD study, attempts to visualise XIP's output in a user-friendly way had been made by De Liddo et al. (2012). The collaboration between The Open University's Knowledge Media Institute and Xerox Research Centre Europe's Parsing & Semantics Group investigated the overlaps and complementarities between the outputs from human analysts and the XIP on 120 documents (De Liddo et al., 2012). The XIP's output of

analysis of these documents was imported into Cohere[10] platform to explore ways of visualising XIP's output together with the human annotations.

These representations provided detailed 'zoomed in' views of the XIP's output, but did not provide effective summary overviews which would allow a reader to grasp the overall quantity and quality of the XIP's analysis, in order to choose where to inspect more closely. Researchers also pointed to future work that proposed tackling the meaning making aspect of the XIP output: "how to make optimal use of both human and machine annotation?" and "how to cope with visual complexity? (new interface, filtering, etc.)?".

In order to address the need to visualise XIP's output in more user-friendly ways to cope with its visual complexity, the pilot study was conducted. The **research question** formulated for the pilot study was:

- **Pilot RQ:** "In what ways can the output of the XIP analysis be visualised in a user-friendly way so that end-users can make sense of this output?"

The next sections describe the pilot study carried out in order to answer this research question.

8.2.2 Visualising the XIP output through the XIP dashboard

8.2.2.1 Dataset

The aim of the pilot study was to explore XIP's results and investigate more user-friendly ways of delivering its raw output. In order to answer the pilot research question and study in what ways the XIP's output can be visualised, a dataset consisting of research papers that potentially included rhetorical elements that XIP looks for was necessary.

[10] http://cohere.open.ac.uk/

During the first year of this PhD study, no research had been carried out regarding XIP's performance on student writing and there was no corpus of undergraduate texts readily available. Since the aim of the pilot study was to investigate ways of visualising the XIP output, it was not necessary that the dataset used for this analysis came from students. Hence, the readily available corpus of The Learning Analytics and Knowledge (LAK) dataset (Taibi & Dietze, 2013) was selected, which is published by the Society for Learning Analytics Research (SoLAR)[11].

As given in section 3.8.2, the XIP was originally developed based on the analysis of experienced researchers' articles in the field of life sciences, whose rhetorical elements and moves established the basis of XIP's automatic analysis of meta-discourse. Therefore, the LAK dataset was chosen, which would potentially carry similar rhetorical moves as it consists of experienced researchers' conference papers and journal articles.

Another reason for selecting the LAK corpus was because it includes the literature for two related research fields: learning analytics and educational data mining. This would a) allow better visual comparison for the XIP output of two different research fields and b) prototype a real-life example of literature review analysis. SoLAR provided access to machine-readable plain text versions of the proceedings of the:

- ACM International Conference on Learning Analytics and Knowledge[12] (LAK) conference (2011 & 2012) and journal special issue[13] (2012).

- International Conference on Educational Data Mining[14] (EDM) conference (2008-12) and journal of EDM[15] (2008-12).

[11] http://www.solaresearch.org/
[12] http://www.solaresearch.org/events/lak/
[13] http://www.ifets.info/index.php?http://www.ifets.info/issues.php?id=56
[14] http://www.educationaldatamining.org/proceedings
[15] http://www.educationaldatamining.org/JEDM/

Taibi and Dietze (2013) stated the challenge that this dataset could exploit: analysis and assessment of the emerging LAK community in terms of topics, people, citations or connections with other fields; and innovative applications to explore, navigate and visualise the dataset (and/or its correlation with other datasets). This challenge could be addressed through a visual analytics dashboard, the XIP dashboard. The initial step was to analyse conference papers and journal articles through the XIP in order to generate a XIP dataset that could be used for the pilot study, which would then be visualised.

8.2.2.2 Design process

The XIP Dashboard, a set of visual analytics modules built on the XIP output using the LAK open dataset as a test corpus, was implemented to initiate an exploration of visual analytics design. The design process consisted of three steps, which are explained further in the following sections:

- reviewing the data visualisation literature in order to motivate the design choices for the visualisations
- paper-prototyping the selected design choices and evaluating them with possible end-users prior to actual implementation of the dashboard
- implementing the XIP dashboard.

8.2.2.2.1 Visualisation

Advances of information visualisation were driven significantly by information retrieval research, whose researchers and practitioners aimed to improve the efficiency and effectiveness of information retrieval (Börner, Chen, & Boyack, 2003). As an academic field of study, information retrieval is defined as: “finding material (usually documents) of an unstructured nature (usually text) that satisfies an information need from within large collections (usually stored on computers)” (Manning, Raghavan, & Sch¨utze, 2008). Accepting the argument that the more a researcher or user knows about her search space,

the more likely it is that her search will be effective. “Many information visualization systems depict the overall semantic structure of a collection of documents” (Börner et al., 2003, p. 7). The XIP in this pilot study was used to process the LAK dataset that consists of a collection of documents. Therefore, the XIP dataset generated through this process could be used within a visualisation environment for information retrieval for making sense of published literature.

“Visualization refers to the design of the visual appearance of data objects and their relationships” (Börner & Polley, 2014, p. 23). Data visualisation is “the representation and presentation of data to facilitate understanding” (A. Kirk, 2016, p. 19). Visual analytics shed light on unexpected and hidden insights and information of data, that may lead to beneficial innovation (Keim, Kohlhammer, Ellis, & Mansmann, 2010). Visual analytics meaningfully render the data and information in a clearer form to end-users who will then comprehend the hidden insights about data more clearly. Humans are the ultimate authority in directing the analysis of data. By combining the strengths of humans and computers, visual analytics sketches a path in understanding of the information presented visually. Visual analytics dashboards are becoming increasingly available in online learning platforms (Siemens *et al*., 2011). The dashboard is the sense-making component, presenting visualised data to assist individuals in making decisions about teaching and learning in learning analytics systems (Siemens *et al*., 2011). As Siemens *et al*. (2011) emphasised the benefits of dashboards to visualise real-time information to learners, educators, and researchers, the aim of the pilot study was to generate a visual analytics dashboard that presented the XIP output in a way that was beneficial for its potential end-users in order to make sense of published literature.

When dealing with visualisations, the viewers will go through a process involving three stages (A. Kirk, 2016, pp. 22-23):

1. *perceiving*, the act of reading the visualisation: ‘what does it show?’

2. *interpreting*, the act of converting these perceived values into some form of meaning: ‘what does it mean?’
3. *comprehending*, the act of reasoning to arrive at a personal reflection: ‘what does it mean to me?”

In order to derive understanding from the visually portrayed data, the presentation should be effective in terms of the design decisions such as selections of charts, applications of interactivity, colour usage and so on (A. Kirk, 2016) that would all facilitate the viewer’s understanding. Well-designed visualisations (Börner & Polley, 2014, p. 23):

- provide an ability to comprehend huge amounts of data
- reduce visual search time
- provide a better understanding of a complex data set
- reveal relations otherwise not noticed
- enable a data set to be seen from several perspectives simultaneously
- facilitate hypothesis formulation.

From the LAK/EDM dataset, the following data information was available: total number of papers for each field, publication year of each paper, and its authors. From the XIP output generated by the analysis of this dataset, the following information became available: total number of rhetorically salient sentences per publication, and distribution of the XIP categories for these sentences per publication. Thus comparison of these based on publication year and field, the emerging LAK community in terms of topics, based on the concept extraction of the XIP and connections with EDM. Therefore, the proposed XIP dashboard would provide its users a better understanding of the large LAK/EDM dataset by noticeably lowering the time of information retrieval, that would otherwise take longer to search and make sense of the research papers in both fields. Additionally, it would reveal relations between both fields regarding the key terms, and rhetorical elements that would potentially facilitate meaning-making.

Data can be represented in many different forms such as line graphs showing trends over time, stacked graphs, or histograms for showing how many observations of a certain value have been made (Börner & Polley, 2014). Ranges of different graph types exist, among them line graphs, bar graphs, and the stacked versions of each, pie charts, scatter plots and bubble graphs are widely used (Börner, 2015, p. 31). Therefore, the widely-used and popular graphs were selected to visualise the XIP dataset that would potentially facilitate meaning-making and comprehending the complex data in an easy and user-friendly way.

- "The most common way to see a trend in data is to plot a variable's change over time on a line chart or bar chart" (Robertson, Fernandez, Fisher, Lee, & Stasko, 2008, p. 1325). Therefore, it was decided to make use of:
    - popular, widely-used and easy to interpret line graphs for visualising the trends of the total number of salient sentences based on their rhetorical type, (XIP category), over publication year for both fields
    - bar graphs to represent the number of papers written based on a specific key term (concepts) in each field (LAK/EDM); and stacked bar graphs to compare the key term, CONCEPT, for both fields together.
- The trend over time can also be shown as an animated bubble chart over time, with the bubbles changing position and size. Hans Rosling used this technology to make presentations at TED (Technology, Entertainment, Design) starting in 2006[16], which evoked a strong and positive reaction from the audience and a favourable response.
    - Therefore, inspired by Rosling's work, it was decided to use bubble graphs on a scattered plot to communicate a) trends of total number of papers published on a specific key term over time, and b) total number of rhetorically salient sentences mentioning this key term and how this

[16] Rosling's TED2006 talk: https://www.ted.com/talks/hans_rosling_shows_the_best_stats_you_ve_ever_seen

changed over time. Additionally, the distributions of rhetorical labels, the XIP categories, per term in specific year were shown through pie-charts.

8.2.2.2.2 Paper prototype evaluation

Prior to the actual software implementation, preliminary design work was done using a multi-screen paper prototype to elicit potential end-users' reactions on the design ideas. The methodology was paper prototyping (Snyder, 2003), which has been used successfully for low-fi storyboarding of user interfaces to elicit rapid feedback on novel interactive applications, and to get user data before coding.

"Paper prototyping is a method of usability testing that is useful for Web sites, Web applications, and conventional software" (Snyder, 2003). Hand-sketched and/or screen shots drafts of the menus, images, pages, windows of the software application are used to simulate how the interface would behave. It is used to test the usability of the software, in which users interact with the paper to accomplish pre-defined realistic tasks, such as clicking paper buttons to open a new window.

Paper prototyping is ideal for finding out whether developers are on the right track or not before writing the code; to revise the interface quickly; and to learn whether or not the paper design meets the expectations of the users. Therefore, prior to the implementation of the XIP dashboard and in order to obtain rapid feedback on the user interface, a multi-screen paper prototype[17] of the design work was tested.

In order to answer the pilot RQ, "In what ways can the output of the XIP analysis be visualised in a user-friendly way so that end-users can make sense of this output?", following the design decisions regarding visualisations as explained in previous section, ideas were sketched on paper as shown in the figure below, to establish whether these make sense to the potential end-users.

---

[17] A five-minute introductory YouTube video of the prototype: http://www.youtube.com/watch?v=HlzN-PviYPM

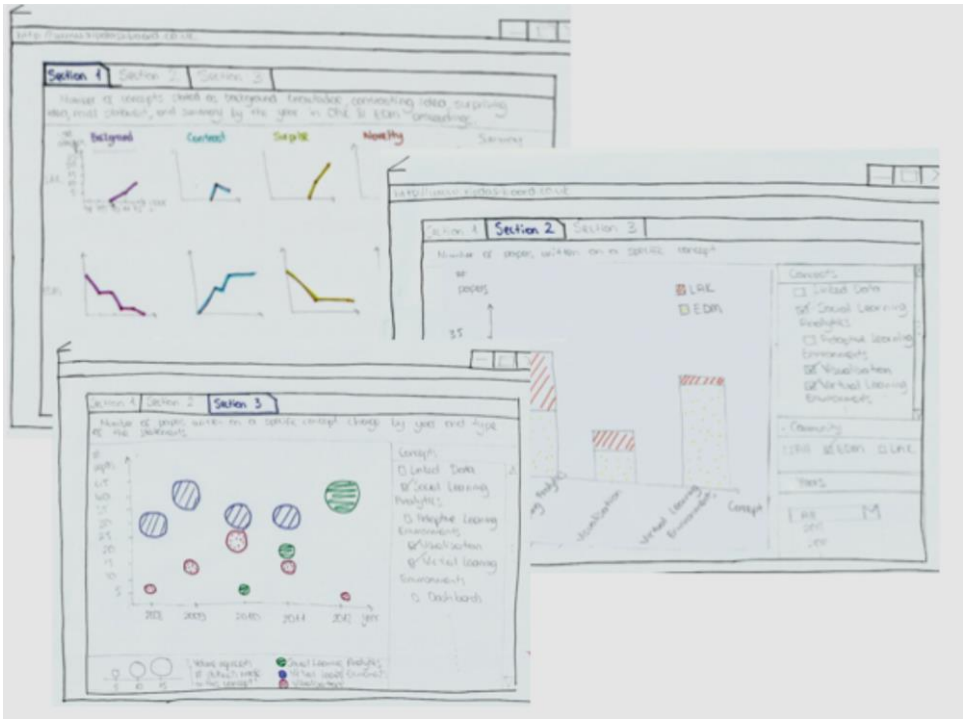


Figure 8.2 XIP Dashboard's paper prototype for user pilot

In order to answer the pilot RQ, six user sessions were conducted with first year PhD students who were just a few months into their literature analyses at The Open University. As the dashboard would be the visual analytics discovery of published literature that would potentially allow the analysis of the literature, first year PhD students were the ideal sample set. This sample set would provide great understanding and evaluation of how these visual analytics elements could be useful and improved, since PhD students are typically working hard on their literature reviews, especially in their first few months of study.

Each session took around 45 minutes and consisted of two phases: testing the user interface and getting opinions about what might need to be changed in the design (Figure 8.1), and discussing how such a tool might address the problems that users were facing in conducting their literature reviews. The specific elements that were trialled in each session were:

1. getting initial and overall reactions to the idea of visualising the literature
2. understanding the usability of the overall design prototype

3. discussing the selection of chart and graph types
4. asking how the design could be improved

The sessions started with a brief introduction to the concept of a machine, the XIP, identifying specific kinds of sentences in papers. An example of raw semantically tagged XIP output was shown (as exemplified in figure 8.2), and even though most of the participants had a computer science background, meaning they knew how to read an XML form output, it took them long time to comprehend all the information even though a briefing was given. The participants then had a guided tour around the interface to explain how the dashboard works (analogous to an instructional movie[18]) and how the XIP's raw output would be shown in a visual way. The participants were then given two tasks, using a think-aloud protocol, and using their fingers as a mouse to 'click' on the screens, which the researcher would then change. All participants managed to navigate the interface and complete the tasks, providing preliminary feedback that the basic information architecture of the prototype was reasonable.

The participants found the raw XIP data very hard to understand and work with, but they found the suggested visualisations relatively straightforward.

The participants found the selection of charts and graphs easy to understand and navigate, and they noted that they particularly liked the idea of using bubble charts to display the data (top sheet in figure 8.2). They made the following suggestions for implementation:

- The participants all suggested merging the line charts (top sheet in figure 8.2) into one plot for each dataset (one for LAK and one for EDM), instead of showing the same thing in ten graphs.

[18] A five-minute introductory YouTube video of the prototype: http://www.youtube.com/watch?v=HlzN-PviYPM

- There was consensus around switching from stacked, multi-variable bars in the charts (central sheet in Figure 8.2), which participants found hard to understand, to a more conventional set of single-variable columns in a bar chart.

At the end of the sessions, all participants expressed enthusiasm about the idea of using such a dashboard for literature analysis which was defined as a time-saving tool by all that could potentially solve the challenges they faced in their literature reviews. They all felt that this could save them time by identifying more effectively the key papers in their research field, and important sentences within these papers along with their rhetorical meaning identified. The positive reactions to this way of visualising the XIP output to make sense of the literature provided encouragement for further evaluations. These discussions took the study forward and after the design refinements based on the suggestions of paper prototype evaluations had been carried out, the XIP Dashboard was implemented, which is explained in the next section.

8.2.2.3 The XIP dashboard

All of the papers within the dataset (66 LAK and 239 EDM papers, 305 in total) were analysed using XIP, extracting 7847 salient sentences and 40163 concepts. The output files, one per paper, were imported into a MySQL database, and the user interface implemented using PHP and JavaScript, making use of Google Chart Tools[19] for the interactive visualisations.

After selecting the document corpus, analysing it using XIP, and storing the output in a relational database, the user interface design was implemented, the design of which had been motivated through the visualisation literature and the preliminary pilot study with students on the paper prototype, as explained in previous sections.

[19] https://developers.google.com/chart/

The XIP dashboard[20] consists of three tabs, each showing different analytical results in different types of charts. The first section consists of two line charts each representing the distribution of rhetorical sentences by year in LAK and EDM. The line charts depict sentences by rhetorical type over time (Figure 8.3). This provides a bird's-eye-view of the distribution of rhetorical moves per year. For instance, both literatures remained stable for most rhetorical types, but there was a clear separation in frequencies between relatively high and rising levels of SUMMARY (shown in green lines) and CONTRAST (shown in red lines) moves compared to the others as shown in Figure 8.3.

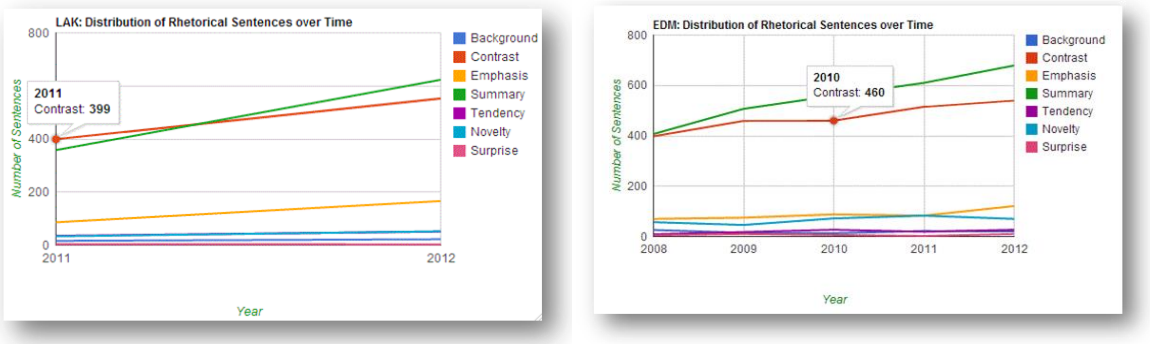


Figure 8.3 Rhetorically classified sentences graphed by frequency and year, for EDM and LAK

The next visualisation of the dashboard (Figure 8.4) permits users to specify a combination of the extracted concepts in which they are interested, in order to see the occurrence of these concepts in papers within all or specified research communities represented in the corpus.

[20] A five-minute introductory YouTube video of the XIP dashboard: https://www.youtube.com/watch?v=AFSHLNFgIVA

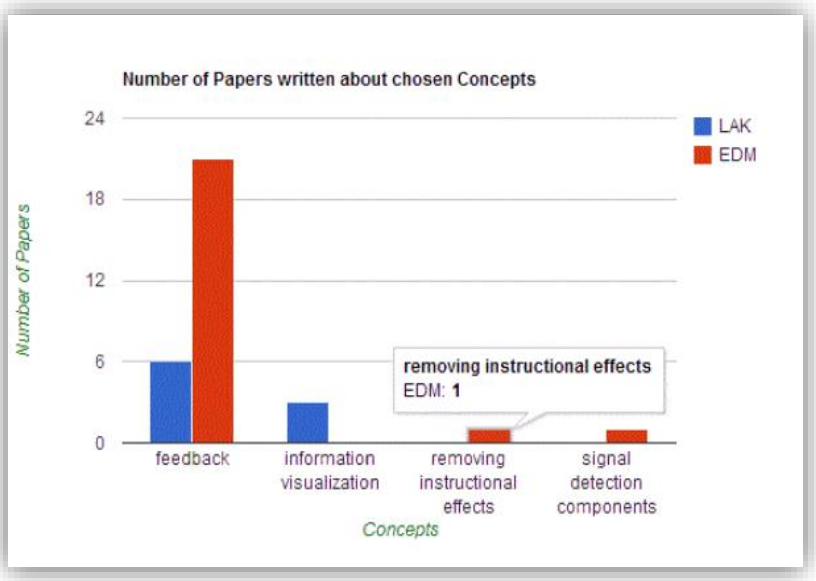


Figure 8.4 Number of papers with rhetorically extracted sentences containing user-selected concepts

Thirdly, a bubble chart displays the occurrence of papers within any or all communities, filtered by user-selected concepts and year of publication (Figure 8.5). As shown by the colour spectrum at the top, saturation represents the 'density' of the concept in the papers, as defined by number of XIP classified sentences in which it occurs (where darker = denser).

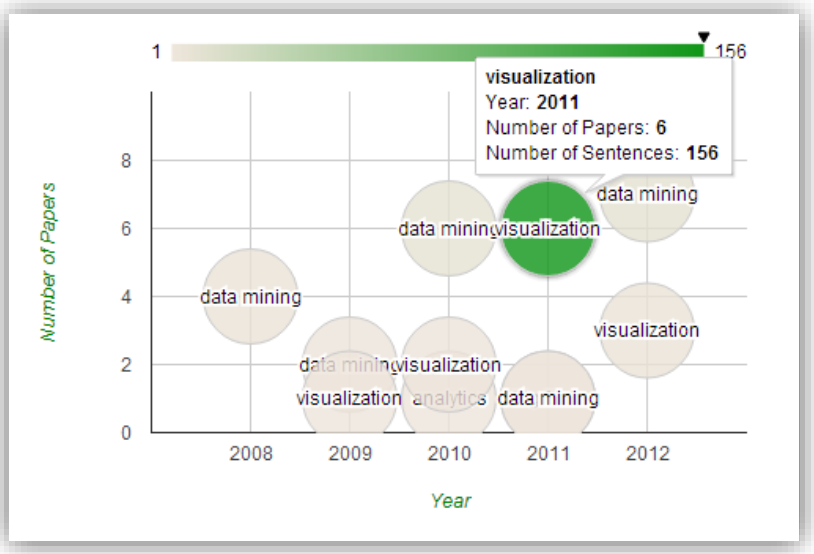


Figure 8.5 Concept 'density' within XIP sentences, by year and number of papers

There is no rhetorical analysis shown at this level of detail, but when a user hovers the mouse over a concept bubble, it displays a pie chart showing the relative distribution of rhetorical types (Figure 8.6).

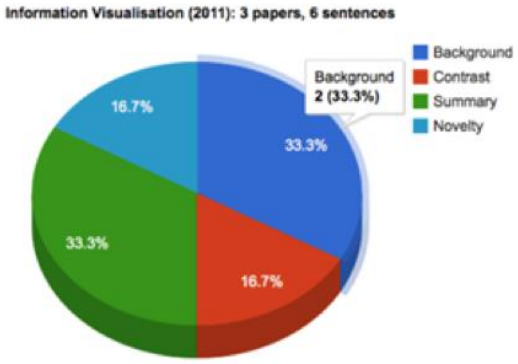


Figure 8.6 Distribution of rhetorical types in the sentences behind a selected concept bubble (Figure 8.5)

When the LAK/EDM dataset was processed through the XIP, the output had multiple inter-related dimensions. This design iteration of a set of visual analytics was a significant advance on working with hundreds of text files, which provided no meaningful support to the end users in seeing any macro-level patterns in a corpus, or navigating between papers sharing common rhetorical moves and concepts.

The next step was to evaluate the potential of this visual analytics dashboard with possible end-users. The proposed working visualisation was first taken to the Learning Analytics and Knowledge (LAK'13) conference as part of an annual data challenge[21] and then demonstrated at the Learning Analytics and Knowledge (LAK'14) conference (Simsek, Buckingham Shum, De Liddo, Ferguson, & Sándor, 2014), where it gained the Best Demo Award[22] on the basis of conference delegates' votes. The aim of the demonstrations was to grasp an understanding of which visualisations work best in this context from the point of view of the user by soliciting opinion from the LAK experts. Based on three hours of informal discussions with 35-40 LAK experts during the demonstrations, there was overall

[21] http://lak.linkededucation.org/
[22] http://kmi.open.ac.uk/news/article/18608

a positive reaction to the design and the idea was noted. No one had negative criticism regarding the visualisations and design decisions. Since most of the delegates had their research papers in the LAK open dataset, they had a chance to see their own work visualised in this way and could make sense of the XIP output; they all were very enthusiastic about the idea. Discussions with the delegates explored ways in which these visualisations could be used and by whom. By gathering expert opinions and taking these forward, the following user scenarios were prepared.

8.2.2.3.1 Possible user scenarios

Discussions with LAK delegates confirmed that this kind of visual analytics dashboard could be used by students or researchers to make sense of the published literature and also by educators who could use a similar dashboard to make sense of their students writing within a cohort (Simsek, Buckingham Shum, Sandor, De Liddo, & Ferguson, 2013).

Student Scenario: preparing for an essay

Jane is a first year Digital Marketing master's student who is enrolled in a Web Analytics module. Part of her assessment requires her to write an essay about How to Get into Big Data Analytics in Online Marketing.

As Jane is new to this research area, her initial step is to carry out a literature review. She's curious to know which community writes most on these topics, so she logs into her university library, which provides subscription access to e-journals and conference papers. She finds the LAK and EDM literature, and then switches to the rhetorical dashboard view, and selects the second tab (Figure 8.4). This allows her to specify a combination of key concepts related to her research, and to see the occurrence of these in papers. She chooses big data and data analytics from the list of concepts and selects 'all communities' to learn which community tend to publish more on these topics.

Once Jane has obtained a general overview about where to look, her next step is getting know the distribution of these numbers by year. Therefore, she moves to the third tab

(Figure 8.5), which displays the occurrence of papers within all communities, filtered by her selection of concepts and year of publication. This shows her when big data and data analytics began to be referenced, in which years this peaked, and the overall trend. She picks a peak year, and wants to find contextualising statements about the background of the topics, for her literature review, so she switches to examining the distribution of rhetorical types (Figure 8.6). She chooses sentences classified as Background topics, and from there, finds the paper listings.

Educator Scenario: assessing essays

Academic tutor, Mrs. Jones is reviewing progress in her advanced level class on Educational Futures. By this stage, the students should be capable of writing coherently-structured essays with a clear thesis, backed by good argumentation, appropriately contextualised to the literature. She brings up the XIP Dashboard and points it to the folder with 45 essays, each 20 pages long. A few seconds later, the visualisations have loaded, and she begins to explore. She can view this year's essays graphed against the preceding years, 2011-2014, giving her a sense of whether there has been an overall change in the use of appropriate concepts, or writing style, but this doesn't seem to be the case. She's a bit annoyed about this, since she's been trying to improve her teaching of scholarly writing. Maybe it's just the students.

Drilling down to individual 2015 essays, she can see that for lower achieving students, the balance of rhetorical moves is quite skewed, with Background and Summary contributions dominating. While using these in the expected introductions and conclusions of their essays, her higher achieving students seem to make stronger, more assertive moves in which they Contrast claims, express Surprise about certain trends (associated with the concepts MOOC and accreditation).

Mrs. Jones finds this so compelling that she applies for approval from the university ethics board to run a pilot with the 2016 cohort to see if dashboard feedback on essay drafts proves useful.

8.2.3 Discussion

In order to help readers to make sense of the published literature and to improve reading skills, the XIP tool was used to analyse the literature within the LAK/EDM dataset automatically and to identify the important sentences within documents together with the rhetorical type of these sentences. The *XIP Dashboard* was designed to provide visual analytics from this XIP output to help its end-users, potentially researchers reviewing literature, to assess the current state of the art in terms of trends, patterns, gaps and connections within the learning analytics and educational data-mining literature.

The pilot study showed that a visual form of the XIP output can be used to provide a scaffold for making sense of the published literature. All student participants who tested the paper prototype noted that as the prototype overviews trends, patterns, gaps and connections within the literature in a user-friendly way, it enables the existing literature to be grasped easily. Once the XIP dashboard was implemented based on the feedback from paper-prototype evaluations, it was taken to expert conference delegates, who voted the dashboard demonstration to be the best one of the LAK'14 conference. This work provided an initial practical contribution to the field which has been taken forward by the University of Technology, Sydney's Connected Intelligence Centre. They have initiated a series of pilots in close collaboration with academics across the faculties and other units to test the potential of the XIP on reflective writing and have implemented an end-user application onto XIP called AWA (Buckingham Shum, Sándor, et al., 2016).

From the discussions, two possible user scenarios emerged. As described in the *Educator Scenario: assessing essays*, such a dashboard could also be transformed and adapted to visualise student essays, and its rhetorical moves used to scaffold writing skills through

visual results so that educators can make sense of their students' essays and provide feedback on rhetorical aspects of their writing. Therefore, on reflection, this scenario found a plausible scenario that this thesis can take further: its aim is to support the educators' essay assessment practices. The pilot study therefore opened up an important question for the main study: "In what ways should the output of the XIP analysis of student essays be delivered to educators so that they can make use of this output to give feedback on student essays?"

8.3 Main Study

As discussed in section 8.2, for the first time the XIP tool's raw output has been visualised in a user-friendly way through a visual analytics dashboard. The pilot study, investigated the ways in which the output of the XIP analysis of writing should be delivered, which led to the fourth subsidiary research question: "In what ways should the output of the XIP analysis of student essays be delivered to educators so that they make use of this output to give feedback on student essays?"

As the fourth subsidiary research question explores the ways in which XIP output can be presented to educators so that they make use of this output to give feedback on student essays, it was necessary to collect user requirements from educators. The design of visualisations starts with a deep understanding of what stakeholders want (Börner & Polley, 2014). XIP connects with education to the degree that there is an overlap in the hallmarks of research articles and the kinds of writing that academics seek to nurture in undergraduate students, and reward through grading criteria (chapter 5). The XIP parser, however, was originally designed to work on peer-reviewed academic research writing by people who were not specifically focused on education or on pedagogic value. Therefore, the user and system requirements of a potential XIP visualisation specifically targeting

educators and aiming to support their assessment practices needed to be elicited, as explained in next section.

8.3.1 The user and system requirements

The ultimate aim of the XIP's educator dashboard is to provide educators with visual analytics reflecting student activity and progress regarding their essays. The visualisation proposed in the pilot study was intended to be adapted for markers whose user and system requirements are discussed below.

8.3.1.1 User requirements

The user requirements for the ultimate educator dashboard were elicited based on the one-to-one interviews with markers as given in chapter 6 and the focus group discussions with writing teachers and researchers, as given in chapter 7. During these sessions, the problems markers had been experiencing regarding assessment and feedback were investigated. The problems of subjectivity in human marking, time limitations, and the need to provide better feedback and examples to ensure students understand their reasoning, became evident in the one-to-one interviews with tutors, the focus group and the literature. Automated support of the XIP embedded in a visual analytics dashboard could potentially solve these problems as follows.

- the subjectivity of human assessment and marking
    - Human markers can disagree with each other and therefore they do not necessarily come to the same conclusion as their peers, which is a reliability concern. The automated output on the other hand is always the same, stays the same and is not subjective. The XIP's educator dashboard could therefore potentially be useful to help educators overcome this problem by offering self-diagnosis to reflect on their assessment and marking.
- the limited time that markers have to assess an essay

    - Benefiting from the XIP output could also potentially help markers to overcome their time problems. Considering it took around 15 minutes for the focus group participants to highlight 13 sentences and an hour for academic tutors to mark four pages of a student assignment, it could be time-efficient to use the automated support, as the XIP analysis took less than a minute for the same piece. Additionally, automated support could potentially help to overcome the subjectivity of human marking.
- the possibility that markers do not necessarily notice that their students are actually making an analytical point, since most of the markers are not linguists
    - Since the XIP has been developed by linguists whose aim was to analyse the rhetorical elements of salient sentences from articles, it could potentially point markers to the analytical points their students are making with its output.

8.3.1.2 System requirements

The system requirements for the potential educator dashboard were elicited based on the problems markers experience as discussed in the one-to-one interviews given in chapter 6 and the focus group discussion given in chapter 7.

During the focus group session, the participants suggested that the XIP should be embedded in a well-designed and developed area, such as course websites or online learning environments. The participants suggested two different dashboards: one for markers and educators, who could input student writing and see what categories the tool assigns, which could then be used to base their feedback on and as a self-help tool to support their communications with their students, i.e. to show examples of good argumentative sentences, and generate discussions; and another one for students who could input their work to revise, edit and critique their own work. Their suggestion has been

taken forward in the main study of this chapter. As Siemens *et al*. (2011) said, "learners and educators require different depths of information in dashboards, based on the context and type of analytics"; considering the scope of this thesis and decisions made based on focus group participants' suggestions, it was decided to design the educator dashboard. Two other points raised by the focus group participants were taken into account as well: the provision of a manual, and the ability to customise the dashboard (chapter 7.6). The focus group participants emphasised that there should be an explanation of what the tool's output and results do/do not mean. The meaning should be transparent so that what it does and what its limits are should be clearly stated. Therefore, any visualisations for the educators should be implemented clearly, explaining what the XIP output means in respect of the students' writing. Since the participants emphasised that what makes good quality student writing is discipline and level specific, there should be an option to customise the tool to adapt it to different levels, disciplines, and modules. This suggested that the ultimate educator's dashboard requires a customisation feature.

8.3.2 Specifications

The user and system requirements described above can be used as a basis for a potential ultimate educator's dashboard that aims to support educators' essay assessment practices using any computational language technology similar to the XIP. The main study proposes an example prototype for such an educator dashboard, hoping that it can inspire further research to implement such a dashboard and evaluate the potential of automatic identification of meta-discourse.

For the implementation of the working prototype, a dataset consisting of student writing was required. The student writing used for generating the XIP analysis to be used in the dashboard came from E000. The reason for selecting the E000 module again was because: 1) the results of the multiple regression study, explained in chapter 5, demonstrated the relation of students' essay marks with the XIP analysis, and produced promising results

about the XIP's performance on E000 students' essays that were in line with tutors' marking rubrics; and 2) E000 markers were interviewed individually as discussed in chapter 6.

The detailed specifications for the working prototype to be built based on the XIP analysis of E000 writing, were identified through an unstructured interview with E000's module chair. During the interviews, the following informative questions were asked to decide on specifications:

- How do educators mark and assess the written assignments?
- What is expected from the educators in their assessment?
- What sort of process do educators follow in essay marking?
- What problems do educators encounter while marking?

The usual process educators go through from the beginning to the end of essay marking for E000 is described as follows. Since the late 1990s, the OU educators use a special software called e-TMA to download all the submitted assignments/essays which are zipped and downloaded into their own computers. Educators then open these documents with any word processor they have and then mark it based on the rubric. Although all E000 markers use a pre-defined marking rubric, they are allowed to choose the way they give feedback. For instance, they can use track changes, or they can give in-line feedback, and they can choose whatever colour scheme, fonts or caps they want to use. It had been mentioned during the interviews that there are inconsistencies between the styles of giving feedback to writing which need to be addressed because all students need to receive consistent and effective feedback. Considering there are more than 60 educators for this module, each choosing their own way of providing feedback, consistency would not be easily achieved. The disadvantage of this could be that if for instance a tutor gives feedback all in capital letters, which is usually perceived as someone shouting in online environments, and using

a red font, which is usually used for pointing to an error, a student might interpret this feedback in a negative way.

Once the markers have finished giving feedback to all the student essays that they are supposed to, they then need to save these and give the final essay grade on another online system called PT3 where educators also have an infinite scrollable text box for their overall comments and feedback which is described as time-consuming by the module chair. As also pointed out in the literature, the one-to-one interviews with markers and the focus group session, markers are pushed for time, so the XIP's educator dashboard needs to save them time rather than adding to their workload. Finally, markers zip all the essays they have assessed and upload them back to the system for students to see it.

From the unstructured interview with the module team chair, the following points have been taken forward for the prototype:

- E000 markers are inconsistent with the feedback styles therefore the XIP's educator dashboard needs to provide consistency to address this issue.
- Markers are pushed for time, so the XIP's educator dashboard needs to save them time rather than adding to their workload. Therefore, it needs to be connected with the systems markers already need to use like e-TMA and PT3.

8.3.3 Prototype

The design of the proposed working prototype was informed by the literature, user and system requirements and specifications, as discussed in the previous section. Based on these, the following design decisions were made.

- **All steps in one**

  The literature review, one-to-one interviews and focus group study with educators suggested that markers have limited time and they are burdened with the overload of marking they are required to do. Therefore, proposing an additional system, a dashboard that educators need to visit besides their usual marking steps, as given in

8.3.2, would take even more time, thus is not plausible. Instead, it was decided that the XIP analysis and output of their students' writing and visual analytics of those need be part of their usual marking steps. That is why the proposed educator's dashboard prototype for E000 markers would allow them to: reach all the essays they need to mark through the dashboard; assess; check the XIP's analysis for these essays; give feedback; and mark them through the same system.

As shown in figure 8.7, when educators logged in to the system, they can see all their students' essays they need to mark. They can select the student whose essay they want to assess. They no longer need to go to e-TMA system to download the essays that they are required to assess as they used to do. They can annotate the students' essay, write comments, give feedback, mark it and save it to the system without going to the PT3 system anymore. Annotating in this way also overcomes the consistency problem that module chair raised in the interview as explained in section 8.3.2, regarding the different feedback styles markers use.

While assessing and giving feedback, markers can also benefit from the XIP's output to reflect on the rhetorical elements just by clicking the 'XIPit' button, shown in amber in the figure below. Using the tool in this way as a self-diagnosis to reflect on their assessment and marking, markers , can potentially check their students' analytical points and rhetorical acts and reflect on the XIP's output. Doing this is up to the markers, it is optional; if they want to see the automated XIP analysis, they can through the 'XIPit' button, but if they do not, they can skip this part. If they do, the XIP's highlights would be shown, displaying the salient sentences that the XIP extracted and when they mouse over they can see how it labelled the sentence. Considering the tensions and power issues raised in previous chapter, in this way markers would not feel they are being forced to use an automated tool but it is their choice whether or not to use it.

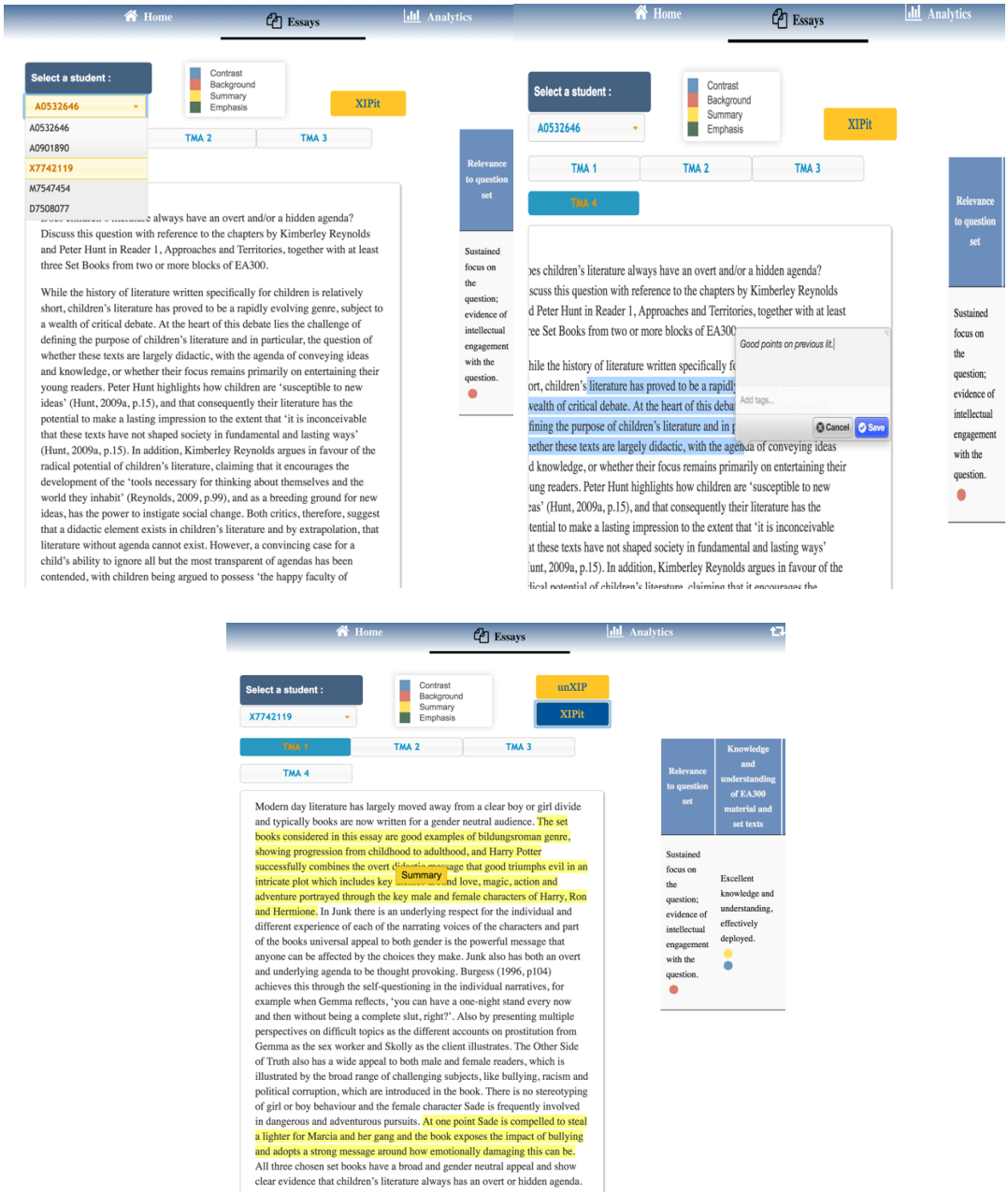


Figure 8.7 The XIP educator dashboard prototype, selection of student essays, annotating and viewing the XIP output

- **Use of XIP categories**

  As discussed in the previous chapter, the XIP needs to unpack the underlying meaning of its categories, and their categories require some changes, additions or revisions. Ultimately, before the potential dashboard is implemented, these concerns should be taken into account. For the prototype, and to exemplify how XIP categories can be used within this context, three XIP categories were chosen: CONTRAST, BACKGROUND, SUMMARY (please refer to chapter 3.8.5 for the

descriptions of the categories). The reason for selecting these categories was because as described in chapter 5, not all of XIP's existing categories were found to have a significant impact on the students' essay mark. The categories TENDENCY, SURPRISE, NOVELTY and OPEN QUESTION, that are found in journal writing of experienced researchers, did not appear necessary for undergraduate students to get better grades.

The ultimate aim was to show the markers the potential of these categories and demonstrate to them how they are linked to their assessment by mapping them to the marking rubrics they use. For the prototype, the XIP categories BACKGROUND (shown in red in figure 8.8) and SUMMARY (shown in yellow in figure 8.8) were linked to the rubric item '*construction of academic argument'*, and CONTRAST (shown in blue in figure 8.8) was linked to '*approach to alternative explanations and arguments'*. These links would allow markers to make sense of XIP's labelling with regard to their marking criteria, and to reflect on their assessment based on XIP's suggestions.

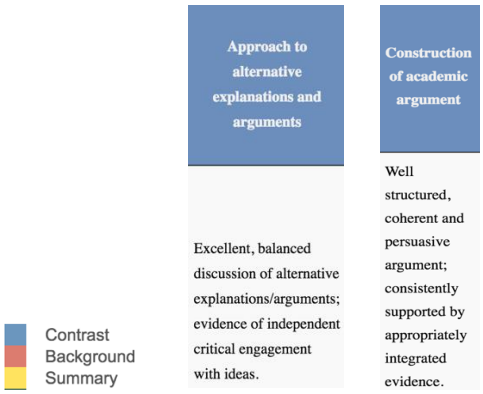


Figure 8.8 The XIP categories mapped with rubric items

- **Analytics**

Considering the time problems that markers face, it was decided not to overload them with graphs that are difficult to interpret. Therefore, based on the reflections

from the pilot study and the literature, it was decided to use the popular, widely-used and easy to interpret graphs: pie charts, bar charts and line graphs.

As shown in figure 8.9, in order to provide educators with a quick overview of the distribution of rhetorical elements in a student's assignment, a pie chart was used, which was motivated from the pilot study. In this example, the pie chart shows that, based on the XIP analysis, student writing mostly had CONTRAST sentences (44.9% of the essay consisted this type of rhetorical act).

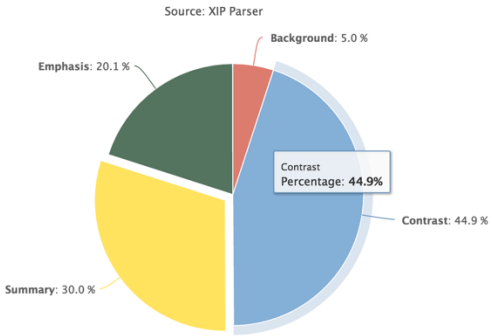


Figure 8.9 Distribution of rhetorical sentence types of one student's assignment

Additionally, as shown in the following figure, tutors were provided with a bar chart combined with a line graph, showing students' overall progress with regard to the XIP output. TMA refers to the tutor marked assignment and in this case the student had three assignments marked by the tutor. The graph shows how this student's grade changed per assignment and what the percentage of salient sentences identified by the XIP was. In this way, the tutor could reflect on how salient sentences affect the essay grade and follow the student's progress.

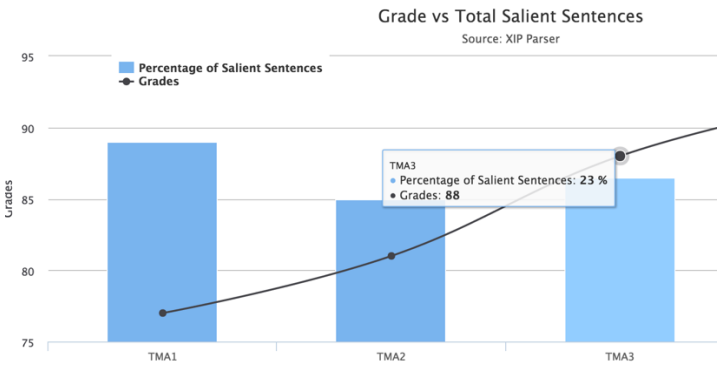


Figure 8.10 Trend of the essay grade compared with the salient sentences

8.3.4 Discussion

The ultimate aim of the XIP's educator dashboard is to provide educators with visual analytics which reflect student activity and progress in respect of their essays. A potential application of such a system by academic tutors would be to use it to inform their judgement, to evidence their judgement and to explain their judgement.

The main study provided an exemplary prototype for this type of dashboard. Following the implementation of the working prototype, it needed to be examined by potential end-users to gain a better understanding of what sort of visualisations would be helpful, and how the current status of the work could be improved and ultimately implemented for the XIP's educator dashboard in future. Therefore, an unstructured interview with E000's module chair was conducted with the aim of understanding whether the design decisions were meaningful and helpful, and how the design decisions could be improved for the benefit of academic tutors.

Providing an environment to academic tutors that includes 'all steps in one' was found to be useful to overcome the consistency and time problems. The graphs and charts were found to be useful and very easy to comprehend. They were helpful in terms of grasping the overall trends quickly. One suggestion was to have the pie chart display on the annotation window, which would have displayed when tutors hit the XIPit button. Since markers are pushed for time, seeing analytics in the marking window could save time.

One significant critique came from the module chair which was related to the explicit use of XIP categories within the marking rubric. Rather than connecting the XIP categories with the rubric items through a colour-coding scheme that would make educators think they are forced to use and adapt new methods into their assessment techniques, mapping should be hidden but done in a more intelligent way. This point made by the module chair is congruent with the 'power and politics' issue, as described in chapter 7.6, theme 3.

In conclusion, the potential visual analytics dashboard intended to be used for providing automated support to academic tutors should not imply that it forces them to use something that would add to their workload. Additionally, anything which might spark their preconceptions and tensions felt towards automated text analysis technologies should be implemented carefully.

9

DISCUSSION AND CONCLUSION

9.1 Introduction

This study has addressed one main research question: 'To what extent can the automated rhetorical parser technology be used to identify indicators of good academic writing in undergraduate student essays and to support educators' feedback processes?' This thesis has investigated whether computational language technologies can automatically identify the attributes of good academic writing in undergraduate student writing and, if this proves possible, how best to feed an actionable analysis back to support educators in their essay assessment processes. In particular, the aim of this thesis has been to understand how automated analysis of meta-discourse in student writing can be used to support tutors' essay assessment practices, to address the gap on how automated essay evaluation can be moved forward to be used in the analysis of higher-order writing skills such as critical argumentative writing, as current literature is lacking in evidence that they can be used for such writing.

The XIP tool has been taken as an exemplar of this type of automated technology to answer this research question. Therefore, its intention has been to increase understanding of how well the XIP tool works for analysing student writing so that it can support educators in providing feedback to students. It has therefore had four aims. The first was to identify in which disciplinary area(s) the XIP tool works best, based on indicators of good student writing, where the given essay mark is used as an indication of the quality. The

second was to build on the previous research by investigating how educators define the attributes of good student writing, and to what degree the XIP tool can identify the presence of these attributes. The third aim was to extend the understanding of how educators and writing researchers judge the quality of student writing and how congruent their views are with the XIP analysis. The fourth aim was to explore what would need to be done to improve the XIP tool and in what ways the XIP analysis of student writing could be used.

Chapters 5 to 8 addressed these aims and provided answers to the main research question by focusing on four subsidiary questions. Chapter 5 examined the XIP analysis of student writing in four disciplinary areas and three undergraduate levels through quantitative analysis. It started with the analysis of the BAWE corpus and validated its results with several OU datasets. This enabled a particular discipline and level to be identified: the third level, education and arts discipline OU dataset, which worked well with the XIP tool. Chapter 6 examined how educators of this OU dataset mark and give feedback to students on their writing, which particular elements they value most, and to what extent these features could be captured by the XIP tool. Chapter 7 extended this study and investigated what educators and writing researchers think about the use of the XIP tool. They were introduced to the XIP for the first time and asked to share their thoughts, worries, doubts and opinions about using the XIP tool to support assessment and feedback. Chapter 8 investigated what improvements, modifications, and alterations should be made to the current form of the XIP tool so that its output can be effectively used by educators when assessing student writing.

This concluding chapter brings together the results reported in the four analysis chapters in order to address the main research question. It begins by identifying the distinctive contributions of this research. It goes on to consider the weakness and limitations of this research and then concludes by showing how this work could be extended and built upon in the future.

9.2 Contributions of this research

This thesis makes the following contributions, which are discussed below:

(i) an investigation of whether automatically identified rhetorical structures can predict essay quality, (section 9.2.1)

(ii) an understanding of how lecturers judge essay quality, (section 9.2.2)

(iii) an investigation of teacher attitudes to automated text analysis, (section 9.2.3)

(iv) and a proposal for a new marking interface design (section 9.2.4).

9.2.1 Can automatically identified rhetorical structures predict essay quality?

The academic writing and literacies literature has shown that academic writing is discipline and level specific. Despite this, there have not been any studies on how automated essay evaluation works in different disciplines and at different levels. This gap, identified in the literature review, has been addressed in this thesis. According to current knowledge, no systematic empirical research exists which evaluates the use of automated text analysis in different disciplinary areas and levels. Therefore, the XIP analysis of a student writing study, given in Chapter 5, provided an empirical contribution to the knowledge. The situatedness of this study and its specific 'discipline and level' elements are where the contributions lie.

The quantitative studies carried out to test the XIP's performance on undergraduate student writing in various disciplines and levels, using the mark awarded as a measure of quality of the writing, suggest that the XIP tool can be used to identify the quality of student writing, especially in soft disciplines such as social sciences, arts and humanities. The current build

of the XIP tool is not helpful in checking undergraduate essays at first level, or in hard knowledge fields.

This study provided an investigation of whether automatically identified rhetorical structures can predict essay quality, which suggested that good quality student writing could be achieved by adopting different meta-discourse, linguistic features, and rhetoric for different disciplines and levels. As emphasised in the literature review, student writers find it difficult to learn how to adapt their writing for different disciplines (Lea & Street, 1998). Thus, based on what has been asked in marking rubrics, students can learn to attend to different XIP categories as partial cues to quality such as BACKGROUND, SUMMARY, CONTRAST and EMPHASIS specifically for the soft knowledge fields. The rest of the XIP categories were found to be out of context for undergraduate writing.

There is the potential for XIP to be used for training students and making them aware of these types of categories in order to improve their writing skills and to help them to get better grades. Practitioners can benefit from this conclusion and once the XIP has been brought up to the level with suggested changes, as given in sections 7.6 and 9.2.3, it can potentially be used for training students to develop good quality writing.

9.2.2 How do lecturers judge essay quality?

This thesis has made a contribution to knowledge through its systematic investigative approach of how to use automated text analysis technology to support educators' assessment practices and feedback processes. The first step was to identify in which disciplinary areas the XIP tool works best, as discussed in chapter 5. The next step was to build on the previous research by investigating how educators define the attributes of good student writing, and to what extent the XIP tool can identify the presence of these attributes. Although the XIP has been found to be working well for the soft knowledge disciplines, it was necessary to identify how academic tutors mark student writing before attempting to use the XIP with students straightaway. This methodological approach,

studying with educators before implementing XIP to automatically analyse student essays is a methodological contribution.

In line with earlier research discussed in the literature (Giltrow *et al.*, 2014; North, 2005), the interviews with markers, as given in Chapter 6, triangulate the research that good quality student writing cannot be defined generically, but is discipline and level specific. This could be the reason that the teaching writing literature has not provided a single gold standard on how to evaluate student writing, but has shared an exemplary assessment criteria provided by Norton *et al.* (2009) that tutors could use when marking students' work. Based on the interviews, this study has produced a well-established, useful resource for stakeholders (researchers, academic tutors, developers of automated text analysis tools for student writing). It consists of a list of five key elements that make for good quality student writing in the soft disciplines, specifically in the field of arts and humanities, as given in Table 9.1 below. Each of these elements has also been related and mapped to the existing XIP categories which is also a contribution of this work.

Table 9.1 Features that make a good quality student writing, which are mapped with XIP categories

1. Criticality	• ability not to parrot surface knowledge but to interrogate the assumptions that lie behind arguments • ability to stand back from one's own perspective and to stand back from the perspective of others and to evaluate them, critique them and synthesise	• CONTRAST • OPEN QUESTION • EMPHASIS • *SYNTHESIS* – (new category required)

2. Argumentation with evidence	• ability to demonstrate criticality through good argumentation which would be evidenced:
	  • by backing up it with materials, literature and readings
	  • by criticising contradicting elements of others' arguments
		• BACKGROUND
		• CONTRAST

3. Voice	• ability to argue by not simply backing up their arguments with references, but taking it to another	• BACKGROUND, CONTRAST, OPEN

	level by challenging the arguments presented by others in order to establish their own argument	QUESTION *and* NOVELTY *all together*
4. Coherence/orientation/ structure	• ability to guide the reader as to the structure of the essay, and to signpost what they are going to write about in the beginning	• SUMMARY
5. Relevant content	• ability to answer the essay question, and write in a relevant manner to the topic	N/A

This contributed to the field as stakeholders can grasp the characteristics of good quality student writing in soft disciplines, specifically in the field of arts and humanities, not only from reviewing the literature but also from the result of this study.

9.2.3 Teacher attitudes to automated text analysis

The literature review showed that human markers and writing researchers have been opposed to the idea of automated text analysis (Elliot and Williamson, 2013; Whithaus in Shermis & Burstein, 2013). There has been an ongoing, inevitable tension between two sides, one side is writing researchers and teachers who are against the overall idea, and who believe that a writing that is so deeply human cannot be analysed by a machine; and the other side is researchers and developers of these tools who support the idea of using the automated text analysis. Since there has yet to be a compromise on either side, this issue is addressed in this thesis, unpacking the underlying factors of this tension further. This study has got to the heart of assumptions, and built the dialogue between both sides; and it comes up with new set of findings.

When educators and writing researchers came together in the focus group, given in Chapter 7, they were initially not inclined to use an automated technology to analyse student writing and did not expect to gain any benefit from it. As expected, the participants' initial reactions were similar to what the literature has pointed out about tension and scepticism. They came to the session with preconceptions about automated

technologies. Their belief was that such technologies had been developed for commercial return, that they can never be as good as human markers, and the aim of using automatic technologies is to automate the marking, not to support assessment processes. However, their ideas have shifted.

Participants were introduced to the XIP tool, not just in theory but in action. They had a chance to see it in action, compare its results with theirs, discuss how it could be improved, and in what way they could use it. Their initial thoughts were concerned with trusting a machine; final thoughts focused on motivating its use. This study revealed that when writing teachers and researchers are introduced to how automated technologies work, they are able to gain a better understanding of such tools' capabilities and limitations, they are reassured that these tools would be used to evaluate the student writing, not score/mark them in order to provide support for the human marker. When they are made part of the process, their opinion shifts.

Yes, these people were suspicious and dubious about the idea, and reluctant to give it a try, thinking it might steal their jobs, that it would never be able to grasp the beauty and subtlety of human language, that potentially it would create tension between human markers and students. But once they had seen it in action and started to compare XIP's results with theirs, they became open to further discussion.

The study showed the key element which made participants open to accepting the idea of using such technology, that is the 'power and politics'. They wanted to feel that they are in control of things and superior over the technology. They wanted to be the 'power' behind such technologies, that should be driven and 'harnessed' by them. Eventually, they wanted to decide how and to what extent they would like to use the automated support. Educators, tutors and markers wanted to be assured that they retain the power themselves in any decision.

Instead of joining the popular narrative condemning writing researchers and teachers who engage with the research about automated text analysis (Whithaus in Shermis & Burstein, 2013), this research signals and suggests an important shift in the ways of introducing and adapting such technologies. It reveals that automated text analysis support should not be provided as the judge, but it should only serve as the lawyer that helps the jury to arrive at the final decision.

9.2.4 Proposal for a new marking interface design

This research has been the first attempt to evaluate the XIP tool, which was originally developed for the meta-discourse analysis of experienced researchers' journal articles, in higher education contexts for the analysis of undergraduate student writing. As discussed in chapter 8, for the first time the XIP tool's raw output has been visualised in a user-friendly way through a dashboard.

The initial pilot study, described in Chapter 8, investigated the ways in which the output of the XIP analysis of writing should be delivered. The proposed working visualisation was demonstrated at the LAK'13 conference, where it gained the Best Demo Award on the basis of conference delegates' votes. Following the pilot study, the main study investigated the ways in which the output of the XIP analysis of student essays should be delivered to educators so that they make use of this output in order to provide feedback on student essays. The main study built on discussions with focus group participants regarding what improvements, modifications and alterations should be made to the XIP tool, so that educators could make use of its output to support their feedback on student writing.

Superficial changes listed in section 7.6 motivated the main study of chapter 8. These changes were about building the XIP tool into a dashboard, supplying it with a manual, and providing educators with the ability to customise it:

- **Built into dashboard**

  The XIP tool needs to be built into a dashboard for markers, who can upload student writing and see what categories the tool assigns, which could then be used to base their feedback on.

- **Supplied with a manual**

  The XIP tool should explain why it has selected certain features. It should be made easier to use with explanations of what its results mean/do not mean. The meaning should be transparent, and an explanation of what it does and what its limits are should be clearly stated so that it can be useful.

- **Ability to customise**

  The tool should allow its users to customise it, and to adapt it to different levels, disciplines and modules.

The visualisation proposed in the pilot study has been adapted for markers, and a prototype for an educator dashboard with a new marking interface has been proposed in the main study of chapter 8 considering the suggested superficial changes based on feedback from the focus group session and based on the user and system requirements which were investigated through unstructured interviews where the participants were selected for their expertise in this area. Therefore, this study has provided a practical contribution.

9.3 Limitations of this research

When the recruitment letter was sent out to academic tutors/markers to take part in one-to-one interviews, the response rate was relatively low. Ideally this study could have been extended through interviews with additional academic tutors. However, this limitation was addressed by holding a focus group session.

The scope of this research was focussed mainly on one specific course, the Open University's E000 module.

The reason for targeting one module in this research was because the results of the multiple regression study, given in Chapter 5 displayed a relation between the essay marks and the XIP analysis, and produced promising results in respect of XIP's performance on E000 students' essays that were in line with tutors' marking rubrics. This research has provided an example systematic investigation for possible future research into the evaluation of automated text analysis across disciplines and levels. The XIP developers have already started to think of ways of customising the XIP tool for different writing styles (Buckingham Shum, Sándor, et al., 2016, in press); and disciplines (Knight et al., Forthcoming). The hope is that this research can inspire further research to evaluate the potential of automatic identification of meta-discourse across disciplines and levels.

9.4 Agenda for future research

This research has suggested further developments for the XIP tool. The following area for future research leading on from this study therefore is to try out an improved version of the XIP tool on a real course with students and academic tutors.

The superficial changes have been dealt within chapter 8 and suggestions of ways in which these changes could be handled have been provided in chapter 8. The profound changes, on the other hand, require time, technical implementation and empirical research by the developers of the XIP tool. These points have already been discussed with the developers of the XIP tool, and responses towards these suggestions are noted as given in the following table.

Table 9.2 Point-by-point discussion with the developers of profound changes suggested for the XIP tool

Requirement for a theoretical underpinning	As discussed in the literature review, there is no explicitly stated linguistic theory behind XIP at the moment. However, although the underpinning theory was not linked to an existing linguistic theory, some theoretical works have links to XIP. Developers at that time did not intentionally ground XIP's modelling of moves in any theory of writing pedagogy or a

	broader learning model, since that was not their background, and it was not the context in which it was developed. However, this clearly does not mean that one cannot argue for, and construct, a mapping between moves that XIP can recognise and a learning model or framework.
Problems due to sentence-level analysis	The developers liked the idea of identifying the constituent concepts in consecutive sentences, and thought it a very good point to improve the tool. How complicated this work would be, is an empirical question. The complexity of this task depends on how the coherence of the target concept, which is ensured by syntactic links within sentences, is captured in sentence sequences. However, the developers of the XIP were positive about the idea and would like to work on it.
Revision of the categories	Focus group participants raised the issue that XIP needs to unpack the underlying meaning of each category. When this point was presented to the developers, it was noted that the categories would always remain 'fuzzy' for the human mind which is an unresolved big issue for natural language processing: XIP Developer: "We try to squeeze things in categories and they are not clear cut. Categories themselves are not clear. It does not pick up the whole meaning." This research has suggested that XIP needs a theoretical underpinning and additional, revised categories that are more pedagogically grounded, and, that one way of grounding them would be to use already existing learning models/frameworks, such as Bloom's Taxonomy (Bloom, 1971). Additionally, the focus group had discussions around whether XIP works lexically, and it that it needs to be based on semantics which was presented to its developers. They clarified that XIP's rules are not lexical but conceptual. However, this new method has never been written down and

	documented; therefore, it is suggested that in order to manipulate the XIP tool, documentation about modelling the XIP tool should be made clear in order to make a real method out of it that people can use.

Section 9.2 showed that this research has important pedagogical implications for the use of automated technologies in both developing undergraduates' writing skills and supporting educators' feedback processes. This research suggests that tutors would benefit from using automated text analysis to support a self-reflection tool. Once brought up to the level, it is possible that it could offer educators a self-diagnosis to reflect on the efficiency of their assessment, feedback and marking.

Once the suggested improvements have been made to the XIP tool, future research initially requires testing to establish whether these improvements are satisfactory. Researchers could investigate the post-reflections through another focus group session. This may then lead to an assessment of the proposed dashboard in real-life settings with a real course, students and educators. As concluded in section 9.2.1, there is the potential for XIP to be used for training students and making them aware of these types of categories in order to improve their writing skills and help them to get better grades. Practitioners can benefit from this and once the XIP tool has been brought up to the level with suggested changes, as given in sections 7.6 and 9.2.4, it can potentially be used for training students to develop good quality writing.

9.5 Concluding remarks

This thesis has investigated whether computational language technologies can automatically identify the attributes of good academic writing in undergraduate student writing and how best to feed back an actionable analysis to support educators in their essay assessment processes. It has achieved its primary aim of increasing the understanding of the extent to which the automated rhetorical parser technology can be used to identify

indicators of good academic writing in undergraduate student essays and to support educators' feedback processes. The XIP tool has been taken as an exemplar of this type of automated technology.

This thesis has proposed that, at the current time, automated text analysis, should not be the sole method of evaluating student writing. Instead, it should be used in combination with human evaluation. It should be recognised that machines do not currently fully understand the language itself, the accuracy of the written material, the content, and the beauty and subtlety of sophisticated argumentation that would be credited by human markers because it flows beautifully. Machines have limited capacity to understand language and literacy; this capacity is mostly dependent on the rules that its developers have written to train them. On the other hand, human language has endless possibilities of creating and forming new sentences each time. Therefore, machines and human markers should complement each other, with the aim of providing better feedback to students. The feedback process of human markers should be supported by machines in order to provide timely, efficient and reliable feedback. Similarly, developers of these machines should improve their tools so that they have a better understanding of what human markers really value and need. Therefore, the relationship between humans and machines should be mutually inclusive rather than exclusive. The developers of such tools should not close the loop without carefully considering the human element, especially markers. This requires resolving the ongoing tensions between the researchers of writing and developers of automated essay evaluation tools. This research has emphasised the importance of the 'decisive power' that academic tutors and markers require to overcome their tension and worry about the use of automated text analysis.

This study is an example of a learning analytics approach that can be followed by the wider LA community should they want to evaluate the potential use of automated text analysis products within learning contexts, for which there is a growing interest. A

significant advance towards understanding the power and effectiveness of automated text analysis in educational contexts has been made, the XIP tool being an exemplar of this type of automated technology. Overall, the aim of this research has never been on grading student writing automatically, but on the potential to identify attributes of good academic writing automatically, so that a computer-aided support for educators could be designed to assist them in monitoring students' progress and in displaying the rhetorical analysis of the essays as formative feedback. This study has played a role in the creation of a dialogue that will form a 'middle ground' between learning and computation, helping members of both communities articulate, in precise terms, the opportunities for pedagogically sound learning analytics.

BIBLIOGRAPHY

Ait Saidi, S. (2015). *Analyse des relations entre concepts scientifiques et structure rhétorique dans des articles de recherche appartenant à un champ disciplinaire des Sciences de l'Education.* (PhD), Université Grenoble Alpes.

Aït-Mokhtar, S., Chanod, J.-P., & Roux, C. (2002). Robustness beyond shallowness: incremental deep parsing. *Natural Language Engineering, 8*(2-3), 121-144.

Andrews, R. (1995). *Teaching and learning argument*. London: Cassell: Continuum International Publishing Group.

Andrews, R. (2003). The end of the essay? *Teaching in Higher Education, 8*(1), 117-128.

Andrews, R. (2009). A case study of argumentation at undergraduate level in History. *Argumentation, 23*(4), 547-558.

Andrews, R. (2010). *Argumentation in Higher Education: Improving practice through theory and research*: Routledge.

Atkinson, D. (1997). A critical approach to critical thinking in TESOL. *Tesol Quarterly, 31*(1), 71-94.

Attali, Y. (2013). Validity and Reliability of Automated Essay Scoring. In M. D. Shermis & J. C. Burstein (Eds.), *Handbook of automated essay evaluation: Current applications and new directions* (pp. 181-199). Oxon: Routledge.

Attali, Y., & Burstein, J. (2006). Automated essay scoring with e-rater® V. 2. *The Journal of Technology, Learning and Assessment, 4*(3).

Attali, Y., Lewis, W., & Steier, M. (2012). Scoring with the computer: Alternative procedures for improving the reliability of holistic essay scoring. *Language Testing, 30*, 125-141.

Attali, Y., & Powers, D. (2008). A developmental writing scale. *ETS Research Report Series, 2008*(1), i-59.

Bacha, N. N. (2010). Teaching the academic argument in a university EFL environment. *Journal of English for Academic Purposes, 9*(3), 229-241.

Baker, S. E., Edwards, R., & Doidge, M. (Eds.). (2012). *How many qualitative interviews is enough?: Expert voices and early career reflections on sampling and cases in qualitative research*. Southampton, GB: National Centre for Research Methods.

Bakhtin, M. M. (1981). The dialogic imagination: Four essays by MM Bakhtin (M. Holquist, Ed.; C. Emerson & M. Holquist, Trans.): Austin: University of Texas Press.

Bazerman, C. (1988). *Shaping written knowledge: The genre and activity of the experimental article in science*: University of Wisconsin Press Madison.

Bean, J. C. (2011). *Engaging ideas: The professor's guide to integrating writing, critical thinking, and active learning in the classroom*: John Wiley & Sons.

Becher, T. (1994). The significance of disciplinary differences. *Studies in Higher education, 19*(2), 151-161.

Bennett, K. (2009). English academic style manuals: A survey. *Journal of English for Academic Purposes, 8*(1), 43-54.

Bennett, R. E. (2011). Automated scoring of constructed-response literacy and mathematics items. *Retrieved April, 14*, 2011.

Berkenkotter, C., & Huckin, T. N. (1993). Rethinking genre from a sociocognitive perspective. *Written communication, 10*(4), 475-509.

Biglan, A. (1973a). The characteristics of subject matter in different academic areas. *Journal of applied psychology, 57*(3), 195.

Biglan, A. (1973b). Relationships between subject matter characteristics and the structure and output of university departments. *Journal of applied psychology, 57*(3), 204.

Bird, S., Klein, E., & Loper, E. (2009). *Natural language processing with Python - Analyzing Text with the Natural Language Toolkit* (1st ed.): O'Reilly.

Birkenstein, C., & Graff, G. (2008). In Teaching Composition,'Formulaic'Is Not a 4-Letter Word. *Chronicle of Higher Education, 4*.

Bizzell, P., Lauer, J. M., Montague, G., Lunsford, A., Emig, J., Maimon, E., . . . O'Connor, F. W. (1982). College composition: Initiation into the academic discourse community: JSTOR.

Bloom, B. S. (1971). *Handbook on formative and summative evaluation of student learning*. New York, NY: McGraw-Hill Book Company.

Bobko, P. (2001). *Correlation and regression. Application for industrial organizational psychology and management* (2nd ed.). Thousand Oaks: SAGE Publications, Inc.

Borg, E. (2003). Discourse community. *ELT journal, 57*(4), 398-400.

Börner, K. (2015). *Atlas of knowledge: anyone can map*: MIT Press.

Börner, K., Chen, C., & Boyack, K. W. (2003). Visualizing knowledge domains. *Annual review of information science and technology, 37*(1), 179-255.

Börner, K., & Polley, D. E. (2014). *Visual insights: A practical guide to making sense of data*: MIT Press.

Boyatzis, R. E. (1998). *Transforming qualitative information: Thematic analysis and code development*: Sage.

Branthwaite, A., Trueman, M., & Hartley, J. (1980). Writing essays: the actions and strategies of students. *The psychology of written communication*, 98-109.

Bridgeman, B. (2013). Human Ratings and Automated Essay Evaluation. In M. D. Shermis & J. Burstein (Eds.), *Handbook of automated essay evaluation: Current applications and new directions* (1 ed., pp. 221-232). Oxon: Routledge.

Bridgeman, B., Trapani, C., & Attali, Y. (2012). Comparison of human and machine scoring of essays: Differences by gender, ethnicity, and country. *Applied Measurement in Education, 25*(1), 27-40.

Brown, M. (2011). Learning analytics: The coming third wave. *EDUCAUSE Learning Initiative Brief, 1*(4).

Buckingham Shum, S. (2008). Cohere: Towards web 2.0 argumentation. *COMMA, 8*, 97-108.

Buckingham Shum, S., De Laat, M., De Liddo, A., Ferguson, R., Kirschner, P. A., Ravenscroft, A., . . . Whitelock, D. (2013). *DCLA13: first international workshop on discourse-centric learning analytics.* Paper presented at the Proceedings of the Third International Conference on Learning Analytics And Knowledge, Leuven, Belgium.

Buckingham Shum, S., & Ferguson, R. (2012). Social Learning Analytics. *Educational technology & society, 15*(3), 3-26.

Buckingham Shum, S., Knight, S., & Littleton, K. (2012). *UNESCO Institute for Information Technologies in Education. Policy Brief.* http://iite.unesco.org/files/anons/19/Foresigh_in_ICT_in_HE_BackgroundDocument.pdf

Buckingham Shum, S., Knight, S., McNamara, D., Allen, L., Bektik, D., & Crossley, S. (2016). *Critical perspectives on writing analytics.* Paper presented at the Proceedings of the Sixth International Conference on Learning Analytics & Knowledge, Edinburgh, UK.

Buckingham Shum, S., Sándor, Á., Goldsmith, R., Wang, X., Bass, R., & McWilliams, M. (2016). *Reflecting on reflective writing analytics: Assessment challenges and iterative evaluation of a prototype tool.* Paper presented at the Proceedings of the Sixth International Conference on Learning Analytics & Knowledge, Edinburgh, UK.

Burstein, J., & Chodorow, M. (1999). *Automated essay scoring for nonnative English speakers.* Paper presented at the Proceedings of a Symposium on Computer Mediated Language Assessment and Evaluation in Natural Language Processing, Maryland, USA.

Burstein, J., & Chodorow, M. (2010). Progress and new directions in technology for automated essay evaluation. In R. Kaplan (Ed.), *The Oxford handbook of applied linguistics* (2nd ed., pp. 487-497). Oxford: Oxford University Press.

Burstein, J., Chodorow, M., & Leacock, C. (2003). *CriterionSM Online Essay Evaluation: An Application for Automated Evaluation of Student Essays.* Paper presented at the IAAI, Acapulco, Mexico.

Castro, F. G., Kellison, J. G., Boyd, S. J., & Kopak, A. (2010). A methodology for conducting integrative mixed methods research and data analyses. *Journal of mixed methods research, 4*(4), 342-360.

Chao, A., Chazdon, R. L., Colwell, R. K., & Shen, T. J. (2005). A new statistical approach for assessing similarity of species composition with incidence and abundance data. *Ecology letters, 8*(2), 148-159.

Chatterjee, S., & Hadi, A. S. (2012). Regression analysis by example. John Wiley & Sons, Inc., Publication.

Chemin, A. (2014). Handwriting vs typing: is the pen still mightier than the keyboard?, Newspaper. *Guardian*. Retrieved from https://www.theguardian.com/science/2014/dec/16/cognitive-benefits-handwriting-decline-typing

Cheville, J. (2004). Automated scoring technologies and the rising influence of error. *The English Journal, 93*(4), 47-52.

Clark, R., & Ivanič, R. (1997). *The politics of writing* (1st ed.). Oxon: Routledge.

Cook, K. C. (2002). Layered literacies: A theoretical frame for technical communication pedagogy. *Technical communication quarterly, 11*(1), 5-29.

Cooper, A. (2012). What is analytics? Definition and essential characteristics. *CETIS Analytics Series, 1*(5), 1-10.

Cox, D. R., & Snell, E. J. (1989). *Analysis of binary data* (Vol. 32): CRC Press.

Crawford, S. L. (2006). Correlation and regression. *Circulation, 114*(19), 2083-2088.

Creswell, J. W. (2012). *Qualitative inquiry and research design: Choosing among five approaches*. Thousand Oaks: Sage publications.

Creswell, J. W. (2013). *Research design: Qualitative, quantitative, and mixed methods approaches* (4 ed.). Thousand Oaks: Sage publications.

Creswell, J. W., Plano Clark, V. L., Gutmann, M. L., & Hanson, W. E. (2003). Advanced mixed methods research designs. In A. T. C. Teddlie (Ed.), *Handbook of mixed methods in social and behavioral research* (pp. 209-240). Thousand Oaks, CA: Sage.

Crossley, S. A., McCarthy, P. M., Salsbury, T., & McNamara, D. S. (2008). *LSA as a measure of coherence in second language natural discourse.* Paper presented at the Proceedings of the 30th annual conference of the Cognitive Science Society.

Dancey, C., & Reidy, J. (2004). *Statistics without math for pshychology: using SPSS for Windows*. United Kingdom: Pearson Education.

Davies, M. W. (2008). 'Not quite right': helping students to make better arguments. *Teaching in Higher Education, 13*(3), 327-340.

Dawson, P. (1998). The rhetoric and bureaucracy of quality management: A totally questionable method? *Personnel Review, 27*(1), 5-19.

De Liddo, A., Buckingham Shum, S., Quinto, I., Bachler, M., & Cannavacciuolo, L. (2011). *Discourse-centric learning analytics.* Paper presented at the Proceedings of the 1st International Conference on Learning Analytics and Knowledge, Banff, Alberta.

De Liddo, A., Sándor, Á., & Buckingham Shum, S. (2012). Contested collective intelligence: Rationale, technologies, and a human-machine annotation study. *Computer Supported Cooperative Work (CSCW), 21*(4-5), 417-448.

de Waard, A., Buckingham Shum, S., Carusi, A., Park, J., Samwald, M., & Sándor, Á. (2009). *Hypotheses, evidence and relationships: The HypER approach for representing scientific knowledge claims.* Paper presented at the Proceedings 8th International Semantic Web Conference, Workshop on Semantic Web Applications in Scientific Discourse, Washington DC.

Deane, P. (2013). On the relation between automated essay scoring and modern views of the writing construct. *Assessing Writing, 18*(1), 7-24.

Denscombe, M. (2008). Communities of practice a research paradigm for the mixed methods approach. *Journal of mixed methods research, 2*(3), 270-283.

Denzin, N. K., & Lincoln, Y. S. (1994). *Handbook of qualitative research*. Thousan Oaks: Sage Publications, Inc.

Denzin, N. K., & Lincoln, Y. S. (1998). The landscape of qualitative research: Theories and issue: Thousand Oaks, CA: Sage Publications.

Dictionary, O. E. *"lexis, n."*: Oxford University Press.

Elliot, N., & Williamson, D. M. (2013). Assessing Writing special issue: Assessing writing with automated scoring systems. *Assessing Writing, 18*(1), 1-6.

Engelhard, G., & Myford, C. M. (2003). Monitoring faculty consultant performance in the advanced placement english literature and composition program with a many-faceted rasch model. *ETS Research Report Series, 2003*(1), i-60.

Ericsson, P. F., & Haswell, R. H. (2006). *Machine scoring of student essays: Truth and consequences.* Logan, UT: Utah State University Press.

Esuli, A., & Sebastiani, F. (2006). *Sentiwordnet: A publicly available lexical resource for opinion mining.* Paper presented at the Proceedings of LREC.

Fairclough, N. (1992). *Discourse and social change*. Cambridge: Polity Press.

Feltrim, V. D., Teufel, S., das Nunes, M. G. V., & Aluísio, S. M. (2006). Argumentative zoning applied to critiquing novices' scientific abstracts *Computing Attitude and Affect in Text: Theory and Applications* (pp. 233-246): Springer.

Ferguson, R., De Liddo, A., Whitelock, D., De Laat, M., & Buckingham Shum, S. (2014). *DCLA14: second international workshop on discourse-centric learning analytics.* Paper presented at the Proceedings of the Fourth International Conference on Learning Analytics And Knowledge, Indianapolis, IN, USA.

Field, A. (2013). *Discovering statistics using IBM SPSS statistics* (4 ed.). Thousand Oaks: Sage.

Foltz, P. W., Kintsch, W., & Landauer, T. K. (1998). The measurement of textual coherence with latent semantic analysis. *Discourse processes, 25*(2-3), 285-307.

Foltz, P. W., Laham, D., & Landauer, T. K. (1999). *Automated essay scoring: Applications to educational technology.* Paper presented at the proceedings of EdMedia, Seattle, USA.

Freedman, A. (1993). Show and tell? The role of explicit teaching in the learning of new genres. *Research in the Teaching of English*, 222-251.

Gabrilovich, E., & Markovitch, S. (2007). *Computing Semantic Relatedness Using Wikipedia-based Explicit Semantic Analysis.* Paper presented at the IJCAI.

Gambier, Y., & Van Doorslaer, L. (2010). *Handbook of translation studies* (Vol. 1): John Benjamins Publishing.

Ganobcsik-Williams, L. (2006). *Teaching academic writing in UK higher education: Theories, practices and models*: Palgrave Macmillan.

Gill, P., Stewart, K., Treasure, E., & Chadwick, B. (2008). Methods of data collection in qualitative research: interviews and focus groups. *Br Dent J, 204*(6), 291-295.

Giltrow, J., Gooding, R., Burgoyne, D., & Sawatsky, M. (2014). *Academic writing: An introduction*: Broadview Press.

Golafshani, N. (2003). Understanding reliability and validity in qualitative research. *The qualitative report, 8*(4), 597-606.

Gorard, S., & Taylor, C. (2004). *Combining methods in educational and social research*: McGraw-Hill Education (UK).

Grabe, W., & Kaplan, R. B. (1996). Theory and practice of writing: An applied linguistic perspective (Applied Linguistics and Language Study): London: Longman.

Haitovsky, Y. (1969). Multicollinearity in regression analysis: Comment. *The Review of economics and statistics*, 486-489.

Haverkamp, B. E., Morrow, S. L., & Ponterotto, J. G. (2005). A time and place for qualitative and mixed methods in counseling psychology research. *Journal of Counseling Psychology, 52*(2), 123.

Herrington, A., & Moran, C. (2001). What happens when machines read our students' writing? *College English, 63*(4), 480-499.

Herrington, A., & Moran, C. (2012). Writing to a machine is not writing at all. In N. Elliot & L. Perelman (Eds.), *Writing assessment in the 21st century: Essays in honor of Edward M. White* (pp. 219-232). New York, NY: Hampton Press.

Hewings, M. (2010). Materials for university essay writing. *English language teaching materials*, 251-278.

Hoepfl, M. C. (1997). Choosing qualitative research: A primer for technology education researchers. *Journal of technology education, 9*(1).

Hounsell, D. (1984). Essay planning and essay writing. *Higher Education Research and Development, 3*(1), 13-31.

Hounsell, D. (1997). Contrasting conceptions of essay-writing. *The experience of learning, 2*, 106-125.

Hyland, K. (1998). *Hedging in scienctific research articles* (Vol. 54): John Benjamins.

Hyland, K. (2002). Authority and invisibility: Authorial identity in academic writing. *Journal of pragmatics, 34*(8), 1091-1112.

Hyland, K. (2004). *Disciplinary discourses: Social interactions in academic writing*: University of Michigan Press.

Hyland, K. (2005). *Metadiscourse*: Wiley Online Library.

Hyland, K., & Bondi, M. (2006). *Academic discourse across disciplines* (Vol. 42): Peter Lang.

Hyland, K., & Tse, P. (2007). Is there an" academic vocabulary"? *TESOL quarterly*, 235-253.

Ivanič, R. (1998). *Writing and identity: The discoursal construction of identity in academic writing* (Vol. 5): John Benjamins Publishing.

Ivanič, R. (2004). Discourses of writing and learning to write. *Language and education, 18*(3), 220-245.

Jaccard, J., Wan, C. K., & Turrisi, R. (1990). The detection and interpretation of interaction effects between continuous variables in multiple regression. *Multivariate Behavioral Research, 25*(4), 467-478.

Johnson, R. B. (1997). Examining the validity structure of qualitative research. *Education, 118*(2), 282.

Kanoksilapatham, B. (2005). Rhetorical structure of biochemistry research articles. *English for specific purposes, 24*(3), 269-292.

Keim, D. A., Kohlhammer, J., Ellis, G., & Mansmann, F. (2010). *Mastering the information age-solving problems with visual analytics*: Florian Mansmann.

Kellogg, R. T., & Raulerson, B. A. (2007). Improving the writing skills of college students. *Psychonomic bulletin & review, 14*(2), 237-242.

Kelly, G. J. (2007). Discourse in science classrooms. *Handbook of research on science education*, 443-469.

King, G., Keohane, R. O., & Verba, S. (1994). *Designing social inquiry: Scientific inference in qualitative research*: Princeton university press.

King, G., & Zeng, L. (2001). Logistic regression in rare events data. *Political analysis, 9*(2), 137-163.

Kirk, A. (2016). *Data Visualisation: A Handbook for Data Driven Design*: SAGE.

Kirk, J., & Miller, M. L. (1986). *Reliability and validity in qualitative research*: Sage.

Knight, S., Buckingham Shum, S., Ryan, P., Sándor, Á., & Wang, X. (Forthcoming). Academic Writing Analytics for Civil Law: Participatory Design Through Academic and Student Engagement. *International Journal of Artificial Intelligence in Education.*

Knight, S., & Littleton, K. (2015). Discourse-centric learning analytics: mapping the terrain. *Journal of Learning Analytics, 2*(1), 185-209.

Kolb, D. A. (1981). Learning styles and disciplinary differences. *The modern American college*, 232-255.

Krueger, R. A., & Casey, M. A. (2000). *Focus Groups. A practical guide for applied research.* (3rd ed.): Thousand Oaks, CA: Sage Publications.

Kuhn, T. S. (2012). *The structure of scientific revolutions*: University of Chicago press.

Landauer, T. K. (2003). Automatic essay assessment. *Assessment in education: Principles, policy & practice, 10*(3), 295-308.

Landauer, T. K., Foltz, P. W., & Laham, D. (1998). An introduction to latent semantic analysis. *Discourse processes, 25*(2-3), 259-284.

Landauer, T. K., Laham, D., & Foltz, P. W. (2003). Automated scoring and annotation of essays with the Intelligent Essay Assessor. *Automated essay scoring: A cross-disciplinary perspective*, 87-112.

Lea, M. (1995). *'I thought I could write until I came here': student writing in higher education.* Paper presented at the Language in a Changing Europe: Papers from the

Annual Meeting of the British Association for Applied Linguistics Held at the University of Salford, September 1993.

Lea, M., & Street, B. V. (1998). Student writing in higher education: An academic literacies approach. *Studies in higher education, 23*(2), 157-172.

Lewin, B., Fine, J., & Young, L. (2005). *Expository discourse*: A&C Black.

Lewis, J. (2009). Redefining qualitative methods: Believability in the fifth moment. *International Journal of Qualitative Methods, 8*(2), 1-14.

Lewis, J., & Ritchie, J. (2003). Generalising from qualitative research. *Qualitative research practice: A guide for social science students and researchers*, 263-286.

Lillis, T., & Turner, J. (2001). Student writing in higher education: contemporary confusion, traditional concerns. *Teaching in Higher Education, 6*(1), 57-68.

Lillis, T. M. (2002). *Student writing: Access, regulation, desire*: Routledge.

Lincoln, Y. S., & Guba, E. G. (1985). Naturalist inquiry. *Beverly Hills, CA: Sage*.

Lisacek, F., Chichester, C., Kaplan, A., & Sandor, Á. (2005). *Discovering paradigm shift patterns in biomedical abstracts: application to neurodegenerative diseases.* Paper presented at the first international symposium on semantic mining in biomedicine.

Loper, E., Yi, S.-T., & Palmer, M. (2007). *Combining lexical resources: mapping between propbank and verbnet.* Paper presented at the Proceedings of the 7th International Workshop on Computational Linguistics, Tilburg, the Netherlands.

Lynch-Biniek, A. (2009). Filling in the Blanks: They Say, I Say, and the Persistence of Formalism. *CEA Forum, 38*(2).

Manning, C. D., Raghavan, P., & Sch¨utze, H. (2008). *Introduction to Information Retrieval.* (1st ed.): Cambridge University Press.

Marcus, M., Kim, G., Marcinkiewicz, M. A., MacIntyre, R., Bies, A., Ferguson, M., . . . Schasberger, B. (1994). *The Penn Treebank: annotating predicate argument structure.* Paper presented at the Proceedings of the workshop on Human Language Technology.

Mayfield, E., & Rosé, C. P. (2013). LightSIDE: Open Source Machine Learning for Text. In M. D. Shermis & J. C. Burstein (Eds.), *Handbook of automated essay evaluation: Current applications and new directions* (pp. 124-135). Oxon: Routledge.

Maynard, A. (1998). Competition and quality: rhetoric and reality. *International Journal for Quality in Health Care, 10*(5), 379-384.

McLellan, E., MacQueen, K. M., & Neidig, J. L. (2003). Beyond the qualitative interview: Data preparation and transcription. *Field methods, 15*(1), 63-84.

McNamara, D. S., Crossley, S. A., & McCarthy, P. M. (2009). Linguistic features of writing quality. *Written Communication.*

McNamara, D. S., Crossley, S. A., & Roscoe, R. (2013). Natural language processing in an intelligent writing strategy tutoring system. *Behavior research methods, 45*(2), 499-515.

McNamara, D. S., & Graesser, A. C. (2012). Coh-Metrix: An automated tool for theoretical and applied natural language processing. *Applied natural language processing and content analysis: Identification, investigation, and resolution. Hershey, PA: IGI Global*.

McNamara, D. S., Graesser, A. C., McCarthy, P. M., & Cai, Z. (2014). *Automated evaluation of text and discourse with Coh-Metrix*: Cambridge University Press.

Mei, W. S. (2006). Creating a contrastive rhetorical stance investigating the strategy of problematization in students' argumentation. *RELC Journal, 37*(3), 329-353.

Mercer, N. (2010). The analysis of classroom talk: Methods and methodologies. *British journal of educational psychology, 80*(1), 1-14.

Miller, G. A. (1995). WordNet: a lexical database for English. *Communications of the ACM, 38*(11), 39-41.

Miller, G. A., Beckwith, R., Fellbaum, C., Gross, D., & Miller, K. J. (1990). Introduction to WordNet: An on-line lexical database. *International journal of lexicography, 3*(4), 235-244.

Mizuta, Y., Korhonen, A., Mullen, T., & Collier, N. (2006). Zone analysis in biology articles as a basis for information extraction. *International journal of medical informatics, 75*(6), 468-487.

Morgan, D. L. (1996). *Focus groups as qualitative research* (Vol. 16): Sage publications.

Mutch, A. (2003). Exploring the practice of feedback to students. *Active learning in higher education, 4*(1), 24-38.

Myers, G. (1992). 'In this paper we report…'': Speech acts and scientific facts. *Journal of Pragmatics, 17*(4), 295-313.

Nagelkerke, N. J. (1991). A note on a general definition of the coefficient of determination. *Biometrika, 78*(3), 691-692.

Nesi, H. (2011). BAWE: an introduction to a new resource. *New trends in corpora and language learning*, 213-228.

Nesi, H., & Gardner, S. (2006). Variation in disciplinary culture: University tutors' views on assessed writing tasks. *British Studies in Applied Linguistics, 21*, 99.

Neumann, R. (2001). Disciplinary differences and university teaching. *Studies in Higher Education, 26*(2), 135-146.

Neumann, R., Parry, S., & Becher, T. (2002). Teaching and learning in their disciplinary contexts: A conceptual analysis. *Studies in higher education, 27*(4), 405-417.

Norris, D., Baer, L., & Offerman, M. (2009). *A national agenda for action analytics.* Paper presented at the National Symposium on Action Analytics.

North, S. (2005). Disciplinary variation in the use of theme in undergraduate essays. *Applied Linguistics, 26*(3), 431-452.

Norton, L., Pitt, E., Elander, J., & Reddy, P. (2009). *Writing Essays at University A Guide For Students, By Students*: London Metropolitan University: Write Now Centre for Teaching and Learning.

Norton, L. S. (1990). Essay-writing: what really counts? *Higher Education, 20*(4), 411-442.

Okada, A., Buckingham Shum, S., & Sherborne, T. (2008). Knowledge cartography. *Software Tools And Mapping Techniques*.

Onwuegbuzie, A. J., & Leech, N. L. (2007). Validity and qualitative research: An oxymoron? *Quality & Quantity, 41*(2), 233-249.

Page, E. B. (1966). The imminence of grading essays by computer. *The Phi Delta Kappan, 47*(5), 238-243.

Page, E. B., & Paulus, D. H. (1968). *The Analysis of Essays by Computer. Final Report.* Retrieved from Storrs, CT: University of Connecticut:

Page, E. B., & Petersen, N. S. (1995). The computer moves into essay grading: Updating the ancient test. *Phi Delta Kappan, 76*(7), 561.

Palmer, M., Gildea, D., & Kingsbury, P. (2005). The proposition bank: An annotated corpus of semantic roles. *Computational Linguistics, 31*(1), 71-106.

Parry, S. (1998). Disciplinary discourse in doctoral theses. *Higher Education, 36*(3), 273-299.

Patton, M. Q. (1990). *Qualitative evaluation and research methods*: SAGE Publications, inc.

Perelman, L. (2012). Construct validity, length, score, and time in holistically graded writing assessments: The case against automated essay scoring (AES). *International advances in writing research: Cultures, places, measures*, 121-131.

Powers, D. E., Burstein, J. C., Chodorow, M., Fowles, M. E., & Kukich, K. (2001). Stumping E-Rater: Challenging the validity of automated essay scoring. *ETS Research Report Series, 2001*(1), i-44.

Ravelli, L., & Ellis, R. A. (2005). *Analysing academic writing: Contextualized frameworks*: A&C Black.

Real, R. (1999). Tables of significant values of Jaccard's index of similarity. *Miscel· lània Zoològica, 22*(1), 29-40.

Real, R., & Vargas, J. M. (1996). The probabilistic basis of Jaccard's index of similarity. *Systematic biology, 45*(3), 380-385.

Rivers, B. A., Whitelock, D., Richardson, J. T., Field, D., & Pulman, S. (2014). *Functional, frustrating and full of potential: learners' experiences of a prototype for automated essay feedback.* Paper presented at the International Computer Assisted Assessment Conference, Zeist, Netherlands.

Robertson, G., Fernandez, R., Fisher, D., Lee, B., & Stasko, J. (2008). Effectiveness of animation in trend visualization. *IEEE Transactions on Visualization and Computer Graphics, 14*(6), 1325-1332.

Roget, P. M. (1911). *Roget's Thesaurus of English Words and Phrases*: TY Crowell Company.

Russell, D. (1995). Activity theory and its implications for writing instruction. *Reconceiving writing, rethinking writing instruction*, 51-78.

Ryvitytė, B. (2003). Research article introductions: variations across disciplines and cultures *KALBOTYRA, The journal of linguistics, 53*(3), 93-100.

Sándor, Á. (2006). Using the author's comments for knowledge discovery. *Semaine de la connaissance, Atelier texte et connaissance, Nantes*.

Sándor, Á., Kaplan, A., & Rondeau, G. (2006). *Discourse and citation analysis with concept-matching.* Paper presented at the International Symposium: Discourse and document (ISDD), Caen, France.

Schwandt, T. A. (2015). *The Sage dictionary of qualitative inquiry*: Sage Publications.

Scott, J. (1999). Looking For the Tidy Mind, Alas. *The New York Times*. Retrieved from http://www.nytimes.com/1999/01/31/weekinreview/looking-for-the-tidy-mind-alas.html?_r=0

Sharples, M. (1999). *How we write: Writing as creative design*: Psychology Press.

Shen, S., & Yao, H. (1999). *The impact of Chinese culture on Chinese learners' use of English writing.* Paper presented at the Proceedings of the 1999 International Symposium on Language Teaching Beijing-Kunming.

Shermis, M. D., & Burstein, J. (2013). *Handbook of automated essay evaluation: Current applications and new directions*. Oxon: Routledge.

Shermis, M. D., & Burstein, J. C. (2003). *Automated essay scoring: A cross-disciplinary perspective*. Mahwah, NJ: Lawrence Erlbaum.

Siemens, G., Gasevic, D., Haythornthwaite, C., Dawson, S., Buckingham Shum, S., Ferguson, R., . . . Baker, R. (2011). *Open Learning Analytics: an integrated & modularized platform. Proposal to design, implement and evaluate an open platform to integrate*

heterogeneous learning analytics techniques. Retrieved from

Siemens, G., & Long, P. (2011). Penetrating the Fog: Analytics in Learning and Education. *EDUCAUSE review*, *46*(5), 30.

Simsek, D., Buckingham Shum, S., De Liddo, A., Ferguson, R., & Sándor, Á. (2014). *Visual analytics of academic writing.* Paper presented at the Proceedings of the Fourth International Conference on Learning Analytics And Knowledge, Indianapolis, India, USA.

Simsek, D., Buckingham Shum, S., Sandor, A., De Liddo, A., & Ferguson, R. (2013). XIP Dashboard: visual analytics from automated rhetor-ical parsing of scientific metadiscourse.

Sinha, R., & Mihalcea, R. (2009). *Combining lexical resources for contextual synonym expansion.* Paper presented at the Proceedings of the International Conference RANLP.

Snyder, C. (2003). *Paper prototyping: The fast and easy way to design and refine user interfaces*: Morgan Kaufmann.

Sommers, N., & Saltz, L. (2004). The novice as expert: Writing the freshman year. *College Composition and Communication*, 124-149.

Soukhanov, A. H. (1999). *Encarta world English dictionary*: St. Martin's Press.

Strauss, A., & Corbin, J. (1990). *Basics of qualitative research* (Vol. 15): Newbury Park, CA: Sage.

Swales, J. (1990). *Genre analysis: English in academic and research settings*. Cambridge, UK: Cambridge University Press.

Swales, J. (2004). *Research genres: Explorations and applications*: Ernst Klett Sprachen.

Symonds, J. E., & Gorard, S. (2008). *The death of mixed methods: research labels and their casualties.* Paper presented at the British Educational Research Association Annual Conference, BERA 2008, Edinburgh, UK.

Taibi, D., & Dietze, S. (2013). *Fostering analytics on learning analytics research: the LAK dataset.* Paper presented at the Proceedings of the LAK Data Challenge at The Third Conference on Learning Analytics and Knowledge, Leuven, Belgium.

Tashakkori, A., & Creswell, J. W. (2007). Editorial: Exploring the nature of research questions in mixed methods research. *Journal of mixed methods research, 1*(3), 207-211.

Teddlie, C., & Tashakkori, A. (2009). *Foundations of mixed methods research: Integrating quantitative and qualitative approaches in the social and behavioral sciences*: Sage Publications Inc.

Teufel, S. (1999). *Argumentative zoning: Information extraction from scientific text.* (Doctoral Dissertation), University of Edinburgh, UK. Retrieved from http://www.cl.cam.ac.uk/~sht25/az.html

Teufel, S. (2006). Argumentative zoning for improved citation indexing *Computing Attitude and Affect in Text: Theory and Applications* (pp. 159-169): Springer.

Teufel, S., & Moens, M. (2002). Summarizing scientific articles: experiments with relevance and rhetorical status. *Computational linguistics, 28*(4), 409-445.

Teufel, S., Siddharthan, A., & Batchelor, C. (2009). *Towards discipline-independent argumentative zoning: evidence from chemistry and computational linguistics.* Paper presented at the Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing, Singapore.

Thonney, T. (2011). Teaching the Conventions of Academic Discourse. *Teaching English in the Two Year College, 38*(4), 347.

Toulmin, S. E. (2003). *The uses of argument*. Cambridge, UK: Cambridge University Press.

Tranmer, M., & Elliot, M. (2008). Binary logistic regression. *Cathie Marsh for Census and Survey Research, Paper, 20*.

Vossen, P. (1998). *EuroWordNet: a multilingual database with lexical semantic networks*: Kluwer Academic Boston.

Vygotsky, L. S. (1988). *The collected works of L.S. Vygotsky: Problems of general psychology, including the volume Thinking and Speech* (N. Minick, Trans. R. W. Rieber & A. S. Carton Eds. 1 ed. Vol. 1). New York: Plenum Press.: Springer Science & Business Media.

Walvoord, B. E. (2014). *Assessing and improving student writing in college: A guide for institutions, general education, departments, and classrooms*. San Francisco: Jossey-Brass.

Wardle, E. (2009). " Mutt Genres" and the Goal of FYC: Can We Help Students Write the Genres of the University? *College Composition and Communication*, 765-789.

Wenger, E. (1998). *Communities of practice: Learning, meaning, and identity*. Cambridge, UK: Cambridge university press.

White, E. M. (1993). Assessing higher-order thinking and communication skills in college graduates through writing. *The Journal of General Education*, 105-122.

Whitelock, D. (2010). Activating assessment for learning: Are we on the way with Web 2.0. In M. J. W. Lee & C. McLoughlin (Eds.), *Web 2.0-Based-E-Learning: Applying Social Informatics for Tertiary Teaching* (Vol. 2, pp. 319-342): IGI Global.

Whitelock, D., Field, D., Pulman, S., Richardson, J. T., & Van Labeke, N. (2014). *Designing and testing visual representations of draft essays for higher education students*. Paper presented at the 2nd International Workshop on Discourse-Centric Learning Analytics, 4th Conference on Learning Analytics and Knowledge, Indianapolis, India, USA.

Whitelock, D., Gilbert, L., Hatzipanagos, S., Watt, S., Zhang, P., Gillary, P., & Recio, A. (2012). *Addressing the challenges of assessment and feedback in higher education: a collaborative effort across three UK universities*. Paper presented at the INTED 2012, Valencia, Spain.

Whitelock, D., Twiner, A., Richardson, J. T., Field, D., & Pulman, S. (2015). *OpenEssayist: a supply and demand learning analytics tool for drafting academic essays.* Paper presented at the Proceedings of the Fifth International Conference on Learning Analytics And Knowledge, Poughkeepsie, NY, USA.

Whitelock, D., & Watt, S. (2007). *Open mentor: Supporting tutors with their feedback to students*. Paper presented at the 11th CAA International Computer Assisted Assessment Conference: Research into e-Assessment, Loughborough, UK.

Williams, J. M. (2007). *Style: Lessons in clarity and grace* (9th ed.). New York, NY: Pearson Longman Publishing Group.

Williams, J. M., & Colomb, G. G. (1993). The case for explicit teaching: Why what you don't know won't help you. *Research in the Teaching of English, 27*(3), 252-264.

Wingate, U. (2012). 'Argument!'helping students understand what essay writing is about. *Journal of English for Academic Purposes, 11*(2), 145-154.

Wolcott, H. F. (2005). *The art of fieldwork* (2nd ed.). Lanham, MD: Rowman Altamira Press.

Xing, M., Wang, J., & Spencer, K. (2008). Raising students' awareness of cross-cultural contrastive rhetoric in English writing via an e-learning course. *Language Learning & Technology, 12*(2), 71-93.

[This page intentionally left blank.]

APPENDIX A: Genre families of the BAWE corpus[23]

Genre Families	Social purpose/ Components/ Genre network	Genres (examples from each family)
Case Study	to gain an understanding of professional practice through the analysis of a single exemplar description of a particular case, often multifaceted, with recommendations or suggestions for future action typically corresponds to professional genres (e.g. in business, medicine, and engineering)	business start-up company report (starts with executive summary) investigation report organisation analysis patient case notes patient report single issue tourism report
Critique	to demonstrate understanding of and the ability to evaluate and / or assess the significance of the object of study includes descriptive account, explanation, and evaluation; often involves tests may correspond to part of a research paper, professional design specification or expert evaluation	academic paper review approach evaluation business environment analysis business / organisation evaluation financial report evaluation interpretation of results legislation evaluation (legal) case report policy evaluation product/ building evaluation programme evaluation project evaluation review of a book/ film/ play/ website system evaluation teaching evaluation

[23] (Nesi, 2011)

Design Specification	to demonstrate the ability to design a product or procedure that could be manufactured or implemented typically includes purpose, component selection, and proposal; may include development and testing of design may correspond to a professional design specification, or to part of a proposal or research report.	application design building design database design game design label design product design system design website design
Empathy writing	to demonstrate understanding and appreciation of the relevance of academic ideas by translating them into a non-academic register, to communicate to a non-specialist readership may be formatted as a letter, newspaper article or similar non-academic genre may correspond to professional writing	expert information for journalist expert advice to industry expert advice to lay person information leaflet job application letter (e.g. reflective letter to a friend; business correspondence) newspaper article
Essay	to develop the ability to construct a coherent argument and develop critical thinking skills may be discussion (issue, pros/cons, final position); exposition (thesis, evidence, restate thesis); factorial (outcome, conditioning factors); challenge (opposition to existing theory); comparison (series of comparative points or arguments); or commentary (series of comments on a text) may correspond to a published academic/specialist paper	challenge commentary comparison discussion exposition factorial

Exercise	to provide practice in key skills (e.g. the ability to interrogate a database, perform complex calculations, or explain technical terms or procedures), and to consolidate knowledge of key concepts data analysis or a series of responses to questions may correspond to part of report or research paper	calculations data analysis mixed (e.g. calculations + essays) short answers stats exercise
Explanation	to demonstrate understanding of the object of study; and the ability to describe and/or assess its significance includes descriptive account, explanation may correspond to a published explanation, or to part of a research paper or professional design specification	business overview concept /job/ legislation overview instrument overview methodology overview organism / disease overview product development overview site/ environment overview species / breed overview substance / phenomenon overview system/ process overview
Literature Survey	to demonstrate familiarity with literature relevant to the focus of study includes summary of literature relevant to the focus of study and varying degrees of critical evaluation may correspond to a published paper or anthology, or to part of a research paper	annotated bibliography anthology literature review notes taken from multiple sources summary book chapter summary series of articles

Methodology Recount	to become familiar with disciplinary procedures and methods, and additionally to record experimental findings describes procedures undertaken by writer may include Introduction, Methods, Results, and Discussion sections, or these functions may be realised iteratively may correspond to a section within a research report or research paper	computer analysis data analysis report experimental report field report forensic report lab report materials selection report (program)development report
Narrative Recount	to develop awareness of motives and/or behaviour in individuals (including self) or organisations fictional or factual recount of events, with optional comments may correspond to published literature, a professional proposal or a report, or to part of a research paper	accident report account of literature search account of website search biography character outline creative writing: short story plot synopsis reflective recount report on disease outbreak urban ethnography
Problem question	to practice applying specific methods in response to simulated professional problems problem (may not be stated in assignment), application of relevant arguments or presentation of possible solution(s) in response to scenario problems or situations may resemble or be based on real legal, engineering, accounting or other professional cases	law problem question logistics simulation medical problem

Proposal	to demonstrate ability to make a case for future action includes purpose, detailed plan, persuasive argumentation may correspond to professional or academic proposals	book proposal building proposal business plan catering plan legislation reform marketing plan policy proposal procedural plan research proposal
Research Report	to demonstrate ability to undertake a complete piece of research including research design, and an appreciation of its significance in the field may include Literature Review, Methods, Findings, Discussion; or may include several 'chapters' relating to the same theme may correspond to a published experimental research paper or topic-based research paper	research paper topic-based dissertation

APPENDIX B: E000 learning outcomes

Knowledge and understanding of:

- Key critical and theoretical debates in the field of children's literature.
- Different theoretical and analytical approaches to children's literature.
- The relationships between the historical development of children's literature and changing conceptions of childhood and of literature.
- The nature of children's literature as a site for ideological indoctrination and struggle.
- The importance of material conditions of production, circulation and consumption of children's literature.
- How literary conventions work within different genres and the role of illustration.
- The significance of change and diversity in children's lives as reflected in children's literature

Cognitive skills:

- Analyse and critically evaluate arguments and evidence, from a variety of sources, relating to particular theoretical and analytical approaches to the study of children's literature.
- Recognise the implicit and explicit assumptions within children's literature about different discourses on childhood.
- Demonstrate the ability to compare and contrast children's books from both thematic and critical perspectives.
- Use appropriate literary, critical and linguistic terminology to describe and discuss specific theories, concepts and evidence.
- Synthesise information and ideas from different sources, including personal experience, in order to reach your own conclusions.

Key skills:

- Read academic and other texts critically, identifying and evaluating positions and arguments.
- Develop research skills, including the ability to garner, sift and organise material and to evaluate its relevance and significance.
- Communicate complex ideas effectively in writing, using the appropriate academic genre and style.
- Construct a coherent argument, supported by evidence and clearly focused on the topic under discussion.
- Continue to develop good practice in the acknowledgement of source material and in the presentation of bibliographies, using appropriate academic conventions.

- Make effective use of the web resources and activities related to the module, and fulfil online activities required for module completion.
- Develop strategies to draw on electronic academic resources in studying children's literature – for example, through accessing online bibliographies, resource sites and children's literature texts.
- Work independently demonstrating initiative, self-organisation and time-management.

Practical and/or professional skills:

- Carry out literary, stylistic and multimodal analysis of children's literature.
- Articulate arguments and develop a synthesis.
- Plan and undertake research, abstracting and synthesising information from a variety of sources.
- Evaluate social, political and ethical issues relevant to children's literature.
- Employ scholarly methods and present material in an orderly and appropriate way.
- Approach problems analytically.

APPENDIX C: E000 marking rubric

	Relevance to question set	**Knowledge and understanding of E000 material and set texts**	**Approach to alternative explanations and arguments**	**Construction of academic argument**	**Clear expression and use of academic conventions**	**Skills of independent study**
Pass 1	Sustained focus on the question; evidence of intellectual engagement with the question.	Excellent knowledge and understanding, effectively deployed.	Excellent, balanced discussion of alternative explanations/arguments; evidence of independent critical engagement with ideas.	Well structured, coherent and persuasive argument; consistently supported by appropriately integrated evidence.	Very clear expression; all sources of evidence appropriately acknowledged and referenced.	High level of motivation; clear evidence of independent engagement with and/or application of ideas.
Pass 2	Clear focus on the question and clear overall direction of answer.	Good knowledge and understanding; appropriate use of a good range of relevant material.	Good discussion of alternative explanation/arguments.	Clearly structured argument; assertions supported by evidence.	Clear expression; sources of evidence generally appropriately acknowledged and referenced.	Evidence of independent engagement with ideas and good motivation to apply insights gained.
Pass 3	Generally clear focus on the question.	Generally draws competently on the most relevant material.	Recognition, and some limited discussion, of alternative explanations/arguments.	Generally clear argument, supported by adequate evidence.	Generally clear expression; adequate acknowledgement and referencing of sources.	Some evidence of independent engagement with ideas; responsive to tutor feedback.
Pass 4	Able to identify and address main issues raised by the question.	Basic understanding of material; some omissions and/or misunderstandings.	Heavily descriptive; little acknowledgement of alternative explanations/arguments	Lines of argument may be clear for short sections but not sustained or developed; supporting evidence thin.	Generally coherent expression; some attempt at acknowledgement and referencing of sources.	Some limited evidence of independent engagement with ideas, and response to tutor feedback.
Bare fail	Some relevant material but fails to address the question consistently or adequately.	Weak understanding or inadequate knowledge of material.	Almost totally descriptive; little acknowledgement of alternative explanations/arguments.	Slight and/or poorly organised argument; inadequately supported by evidence.	Not well expressed; poor acknowledgement of sources.	Weak evidence of independent engagement with ideas; poor response to tutor feedback.
Fail	Frequently loses sight of the question and/or does not demonstrate understanding of the question.	Very little from module or fundamental misunderstandings.	Virtually no acknowledgement of alternative explanations/arguments.	Very weakly argued and evidenced.	Serious deficiencies in expression and/or acknowledgement of sources.	Little evidence of independent engagement with ideas, or response to tutor feedback.
Bad fail	Material mostly irrelevant to the question.	No, or very little, evidence of exposure to E000 material.	No acknowledgement of alternative explanations/arguments.	Little attempt at argument; no supporting evidence for claims.	Little written or not in full sentences; no acknowledgement of sources.	No evidence of independent engagement with ideas; unresponsive to tutor feedback.

APPENDIX D: E000 marking guidelines

Relevance to question

The EMA questions are intentionally broad and there is no set prescription for interpreting them. For both options, students should not only discuss and compare their chosen texts but should also address the broader question. For Option 1 this means including some consideration of how children's literature intersects with cultural change, and for Option 2 some discussion of the history and tradition of children's literature.

Knowledge and understanding of course material and set texts

Students may use any of the Set Books for either of the options, without exception, and may use any editions of the Set Books or A Monster Calls. They should not gain credit for plot summaries.
Students should lose marks if they do not draw on the minimum required number of Set Books, three, for Option 1 but there is no requirement to devote equal space to them. For each option, students can choose to refer to more than the minimum specified number of Set Books. Those students who limit their essays to the specified number of Set Books should be able to achieve as high a grade as any others, and so markers should neither deduct nor award credit solely for the use of more Set Books. Students may write about the same Set Books they chose for previous assignments but should not re-use material, such as sections of previous assignments whether in the original or in an edited form. If such re-use occurs it may be noticeable in relation to the 'relevance to question' criterion but markers are not expected to investigate further.
Students who do not also refer to E000 module materials apart from the Set Books, or who do not use material from more than one block, should not pass as they will not have demonstrated sufficient knowledge and understanding of the module.

Approach to alternative explanations and arguments

Students are expected to engage critically with their own selection of module materials. They should gain credit for their selection of materials and quality of critical engagement.

Construction of academic argument

Markers can give credit for students' rationales for their selection of Set Books, outlines of the framework for their discussions, use of examples, and presentations of coherent and balanced arguments.

Clear expression and use of academic conventions

The Harvard system is recommended but is not compulsory. Students should not lose marks for using a different system if it is used consistently.
Markers should distinguish between inaccurate or insufficient referencing, which can be addressed through deducting some marks against the criteria for expression and academic conventions, and suspected actual plagiarism which should be drawn to the attention of the EAB on the marking form. All student EMA scripts will be centrally scrutinised for plagiarism using Turnitin and Copycatch software.

Skills of independent study

The EMA in its entirety assesses this skill. It is primarily about students' independent engagement with ideas in the module, rather than external research. External sources may be used if relevant, provided that they are appropriately referenced, and students can gain some credit providing this use goes beyond listing additional sources and reproducing material from them. The important question is whether any external material is deployed in relevant and appropriate ways and demonstrates engagement with ideas. It is not essential to use additional resources, and priority should be given to E000 material.

APPENDIX E: E000 tutor recruitment letter

Would you like to help a PhD student and receive a £30 voucher from M&S?

This post is from OU PhD student Duygu Simsek who needs five E000 Associate Lecturers to help her understand how they mark student essays. Thanks for taking the time to read this request.
Best wishes,

_
Dear Associate Lecturers,
I am a 3rd year PhD student at the Open University (OU), investigating whether computational language technologies can automatically identify attributes of good student writing, and how best to feed back the results of such discourse analysis to support learners and educators. My research does not aim to automate the marking process but does aim to support educators as well as students to give and receive better feedback.
I need your assistance to help me in developing better ways of supporting student writing. I am contacting you to ask if you would be willing to take part in this research. Participation in this research would involve you:

1) Meeting with me for max. 1.5 hours either on campus, if you are based in Milton Keynes, or via Skype/Lync.
2) Assessing one E000 student essay during that meeting
3) Highlighting key sentences in this essay which you think demonstrate the competencies that help you to give a high mark
4) Discussing your reasoning behind this with me during that meeting

This study has been approved by the OU's Student Research Project Panel. All data will be treated in strictest confidence. The only people who would see information relating to your part in the study would be myself and my supervisors. Personal information will be stored on a computer and storage will comply with the 1998 Data Protection Act. Any data used to illustrate research reports will be anonymised so that individuals cannot be identified.
If you interest in taking part in this research, you will be free to withdraw at any time up to a week after the meeting, in which case any data relating to you will be destroyed at your request. If you would like to know more information about the research project please do not hesitate to email me.
Please e-mail me by close of play on 10 December, if you'd like to participate in this study so that I can book your slots.
I would really appreciate your support in this endeavour.
Many thanks for taking the time to read this post.
Yours sincerely

Duygu Simsek
PhD Research Student

E-mail: duygu.simsek@open.ac.uk

APPENDIX F: E000 semi-structured interview questions

First of all, I'd like to THANK YOU for agreeing to take part in this study; I really appreciate your time. I'll be audio recording our session today; would that be alright? My aim today is to understand how you assess your students' writing, which aspects of students' essays you value most, what matters to you and what makes a good essay for you. Our interview consists 3 sections; section 1 is general introductory section where we briefly discuss your background and views on assessment, good student writing; section 2 is the main highlighting exercise and section 3 is the follow-up to this exercise.

Section 1:
I will start asking some general questions before passing onto the highlighting exercise in order to understand your views on assessing a student writing.
Demographic Questions:

1. When have you started working as an AL?
2. For how many years you have been teaching E000 groups?
3. What is your own discipline? Which subject area do you come from?

What makes a Good Essay?

4. How would you define a good piece of student writing?
5. What are the key components of a successful essay? The key elements in students writing? (e.g. clarity, argument, structure or syntax, punctuation, layout well-argued, well-structured etc.)
    a. Could you explicate further the terms you used, well-argued, critically analyse, evaluate, synthesise?
    b. What a well-written argument looks like in a written assignment?
6. I know you use marking guidelines, rubrics while assessing your students' essays but which aspect do you care most, what is it you are always looking for?
7. There are 6 aspects in the E000 marking rubric, which two is the most important for you? (relevance to question, knowledge and understanding, approach to alternative explanations and arguments, construction of argument, clear expression and use of academic conventions, and skills of independent study)
8. Does discipline have an effect on what makes a good essay? (If you have marked essays from various disciplines or let's say from various modules, what could you say about the differences between students' essay?
9. Do you believe that your own background, discipline where you come from, influence what you think a good essay is? (Does your own disciplinary history have an influence on what you think a good essay is?)
10. We talked about different disciplines, if we move to different levels of students for same disciplines; what would you expect to see in the writing of a Level 1 student, and how would you expect that to change in the case of a Level 3 student? (such as would you expect level 1 students to give background information most and summarise the literature but level 3 synthesise that information and produce their own judgements, arguments)

Section 2:
OK. Now we completed section 1; so we can move to our main section which is highlighting exercise.
Here I bring one of the students' essays (EMA) from E000-13J representation. I also brought some highlighters and what I kindly request from you to do is *highlighting the key sentences that you think demonstrate the competencies that help you to give a high mark, sentences that has a positive effect on awarding student a good mark*.
Also while highlighting I want you to think out loud. You can read aloud the sentences and please do share your reasoning with me while questioning what your student has done (such as oh well this sentence is providing some information about existing literature which is good).
What I want to see here is how you approach an essay while marking and what makes an essay good, which sentences affect the grade you give.

Section 3:

11. Tell me for 3 mins. what was good about the essay and what was bad in this essay?
12. As you know this essay has already marked, so what do you think is the awarded mark? (or a grade range)
13. This essay was given 92 do you agree? What do you think was the positive affect on this etc.? Do you think that grade aligns what you find in the structure do you think other things might affect this grade?

14. What difficulties you encounter when assessing your students' essays? (e.g. time, deciding what matters most?)
15. What kind of support mechanism would help your assessment?
16. What would you like to see in advance? Such as would you like to see an already highlighted student essay? A document that shows the key sentences with an explanation of what the sentence is such as background info about existing literature, contrasting his own ideas with the lit., summarising the key points?

Would you like to be contacted further regarding this research?

APPENDIX G: transcription guidelines

List of Acronyms and interviewees	• DB: Duygu Bektik (interviewer and researcher) • Interviewees are anonymised as **Tutor 1-7** for individual interviews, chapter 6, and as **Participant 1-6** for focus group, chapter 7.
Dashes	• The em-dash (—) is used to signify an interruption. • A short dash (-) is used to indicate a change of course mid-sentence • A short dash is also used when the speaker stops short and then repeats him/herself, or abandons the utterance completely, restarting with a new sentence. • The thought is enclosed in dashes (- -) if the sentence veers again or goes back to its original course.
Filler words, filled pauses	• A full verbatim transcription approach was followed for the focus group, chapter 7; therefore, filler words like 'um, err' are only used for that section. No special mark-up is used for those words. • Since an intelligent verbatim transcription approach was followed for one-to-one interviews, chapter 6, filler words are omitted from the transcriptions.
Ellipses	An ellipsis is three dots, each separated by a space (. . .) in parenthesis is used to express an omission. For instance, if an interviewee uses a full name, this is omitted for anonymity, or anything that would make the interviewee identifiable is omitted and ellipses used instead.
Quotes	Single quote '' is used for quoted dialogue, such as when interviewee reads a passage from a student text.
Non–Verbals	• For the transcription of the focus group session, chapter 7, non-verbal communication such as laughter, gestures, etc., double parenthesis and italicized text is used to indicate non-verbal communication. Example: ((*Laughs*)) • For pauses in the recording for a meal or other occurrence, the word "break" (all caps) is placed in brackets to indicate the pause. Example: [BREAK]
Unclear, inaudible or unintelligible speech	When transcribing a difficult, indecipherable, an entirely unintelligible passage or words, it is indicated in the text by placing brackets around the word "inaudible" followed by exact audio timing and question mark. Example: [0:35:44.1?inaudible]