Towards Mining Sequences and Dispersion of Rhetorical Moves in Student Written Texts
Simon Knight, Roberto Martinez-Maldonado, Andrew Gibson, and Simon Buckingham Shum
Connected Intelligence Centre, University of Technology Sydney, Sydney, Australia

ABSTRACT
There is an increasing interest in the analysis of both student’s writing and the temporal aspects of learning data. The analysis of higher-level learning features in writing contexts requires analyses of data that could be characterised in terms of the sequences and processes of textual features present. This paper (1) discusses the extant literature on sequential and process analyses of writing; and, based on this and our own first-hand experience on sequential analysis, (2) proposes a number of approaches to both pre-process and analyse sequences in whole-texts. We illustrate how the approaches could be applied to examples drawn from our own datasets of ‘rhetorical moves’ in written texts, and the potential each approach holds for providing insight into that data. Work is in progress to apply this model to provide empirical insights. Although, similar sequence or process mining techniques have not yet been applied to student writing, techniques applied to event data could readily be operationalised to uncover patterns in texts.

CCS Concepts
• Applied computing – E-learning • Computing methodologies – Natural language processing

Keywords
Learning analytics, writing analytics, temporal analysis, sequence mining, process mining, text mining, rhetorical moves, academic writing

1. INTRODUCTION
There is an increasing interest in the analysis of both student’s writing [3] and the temporal aspects of learning data [5, 25]. In order for text-based analytics to support higher level learning – such as the sharing of ideas together, or the communication and critical analysis of those ideas – analysis of data with a ‘temporal’ or sequential character is necessary [24, 25]. In the context of writing, this analysis might include data regarding particular writing processes, the behaviours they involve, and the outputs they produce (for example, note-taking, drafting, copy-editing); or of sequences, both of events (for example, the co-occurrence of inserting a citation and inserting language from the referred to paper), or of linguistic information within a text (for example, argumentative moves that recur in consistent sequential forms). Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from Permissions@acm.org.

These analyses are further complicated by the potential for them to be multi-modal (for example, across online and paper-based documents), and multi-agent (for example, including interaction with peers, tutors, and others). This paper first discusses extant literature on what we characterise as ‘temporal writing analytics’ (§2), highlighting cohesion (§2.1) and rhetorical move (§2.2) analyses. We illustrate this by our developing work on a corpus annotated with ‘rhetorical moves’ and their analysis [16]. Secondly, based on literature on temporal analysis of writing, developing analysis and sequence pattern mining, and our own first-hand experience of sequential analysis, we propose four approaches to pre-process and four approaches to mine frequent sequences and dispersion of rhetorical moves in authentic student writing, highlighting their potential for analysis of features of that writing whole-texts (§3-4).

2. SEQUENCE AND PROCESS ANALYSES OF STUDENT WRITING
A small body of work in the literature has explored temporal sequence and process features of student writing. For example, to study writing processes, analysis has been conducted on the editing features in Google Docs to investigate the revising and editing processes students undertake in writing tasks [4, 27, 41], with similar work in an automated writing tutor tool [37]. In other work, researchers have explored the ways that topic modelling techniques, and social network analyses, can be used to investigate the emergence, divergence, and convergence of ‘voices’ throughout a document [10], describing an essay’s ‘flow’. Similarly, argument mining refers to “the automatic identification of the argumentative structure contained within a piece of natural language text” [26]. The presence and combination of specific elements of arguments (e.g. premises, argumentative moves, claims) are fundamental to definitions of “argument” and its application to computational modelling of argumentation [e.g., 15, 45], thus lending itself to sequence analysis for argument. Consequently, the potential of sequence analysis is also being investigated in terms of identification of identity, argument structures and arguments in use through the analysis of multi-argument constituent parts (e.g., 28).

Obvious syntactic dependencies, represented as sequences of parts of speech in essays, can also be used to infer structures – for example, causal relations [18], or term-definition pairs [14] – present in a text, with the additional analysis of temporal references in a text supporting identification of question-answer pairs [2], and types of discourse [38]. Association rule and sequence mining approaches have been similarly used to identify relationships between constructs in a text [for example, 1, 20, 35], and to detect erroneous sentences [42]. Linguistic research has used these approaches to investigate the ways that language is structured, and used in everyday contexts.

2.1 Text Cohesion
In research specifically on student writing producing a limited body of

Field Code Changed
Field Code Changed
Field Code Changed
Field Code Changed
Field Code Changed
Field Code Changed
those texts using natural language processing (NLP) techniques. Using linguistic features related to textual cohesion and lexical sophistication, the position of a paragraph – introductory, body, or conclusion – can be predicted with a similar degree of accuracy to human judgements (65% against 66% accuracy), indicating that different sections of texts tend to contain particular features [11]. Moreover, higher quality paragraphs were easier to rate (presumably because they best exemplified the rhetorical structure of the intended paragraph type), and paragraph-position specific information regarding ‘body’ paragraphs may inform pedagogic strategies [32].

In addition to textual indicators being used to identify paragraph-position, indicators can be constructed with small but significant predictive value for ratings of introductory ($r^2 = 0.25$), body ($r^2 = 0.10$), and concluding ($r^2 = 0.11$) paragraph-quality [36]. This analysis indicated that good introductions tended to be longer, using more frequently occurring words, plus indicators of temporal cohesion (consistent tense, aspect, etc.) and causal cohesion (i.e. indicators of how events and actions are connected).

Body paragraphs, then, displayed deep vocabulary alongside locational nouns, more varied sentence structures, and an internal cohesion. Finally, conclusions “express more specific ideas using accessible, yet varied syntax. This pattern is consistent with the rhetorical goal of concluding an essay with a straightforward summary of one’s idea that provides the reader with a ‘big picture’ understanding” [36].

### 2.2 The Rhetorical Structure of a Text

The work of Swales [36] hypothesised that, for example, relationships identified in introductory paragraphs indicate authors’ previewing of arguments and provision of reasons for their position. However, the measures used provide limited insight into rhetorical structure. Indeed, building on this work, Crossley et al. [12] further analysed the use of key n-grams (unigrams, bigrams, and trigrams) in poor and high quality introductory, body, and concluding paragraphs, by grouping the key n-grams into clusters of “rhetorical, grammatical, syntactic, and cohesion features” [12]. Across all three paragraph positions, rhetorical features were identified as most important, indicating the potential of analysing rhetorical n-grams within paragraph-contexts to inform automated approaches to assessing whole-essay quality [12].

An alternative approach to understanding sub-patterns within texts has emerged from work informed by Swales’ analysis of rhetorical moves in academic writing [43]. In research paper introductions, these moves mark the: introduction of a topic and its background; raising questions, contrasts or the need to extend the topic; and establishing the contribution or novelty of the given text (its niche or the gap it occupies). Swales thus offers a template for thinking about the kinds of linguistic patterns (rhetorical moves) that occur in texts, where in texts they might appear and in what kind of sequence. Specifically, we would anticipate that texts make ‘moves’ between: (1) introducing topic background, (2) establishing a space within that topic through its critique, contrast, discussion of its potential for extension, etc.; and (3) filling that space by taking a particular position or stance with regard to it. These moves might be dispersed both within a particular section of text (i.e., we would expect more ‘1’ moves earlier, and more ‘3’ moves later in an introduction), and within particular sub-section-sequences such as individual paragraphs. Thus, we might expect paragraphs to introduce background (move 1), evaluate that information (move 2), and point to the potential for future research (move 3). In related work on ‘Argumentative Zoning’, location is used as a feature in order to determine the rhetorical move being made in a scientific text [17, 44].

One tool that has investigated automated approaches to the detection of these moves at a sentence level, and specifically in a higher education context, is the Intelligent Academic Discourse Evaluator (IADE) [7], research on which has demonstrated the variation in the weight of particular rhetorical moves in different disciplinary texts. Of course, analysis of rhetorical moves above the sentence level (or, moves as comprising many sentences) is important too [6, 13], and indeed a tool based on these rhetorical moves has been developed to give feedback on the moves within abstracts [13], with a newer tool (the Research Writing Tool – RWT) developed to support students in reflecting on the rhetorical moves that should be present in each section of their text, in their disciplinary context [8, 9].

#### 2.3 Summary: Developing Analytic Approaches

Across the body of work reviewed above that explores this issue, analysis has indicated that sections of a text (for example, introduction, body, conclusion) contain different linguistic features, and that the presence of these features may be an indicator of the appropriateness (or quality) of those sections. That is, good introductions tend to have particular characteristics as compared to other text sections and poorer quality introductions. Across this work, the rhetorical structure of a text has emerged as a particularly salient feature. Expert analyses of student writing have investigated the weight of particular rhetorical moves, or textual features, in sections of a text. These analyses have been related to human assessments of the ‘type’ of text observed (i.e., whether it was introduction, body, or concluding text); disciplinary genre; and – in a few cases – the quality of the text observed. This analysis has tended to focus on describing the nature of particular text sections, but less on relating these to assessment criteria judgements, or feedback to students. Moreover, expert prior analyses have not, to our knowledge, extended to the potential of sequence and process mining approaches for understanding textual data.

### 3. SEQUENCE AND PROCESS MINING ON STUDENT DATA

### Sequential and Process Mining on Student’s Data

Sequential mining and process mining are techniques that have been used to identify patterns in educational datasets by considering the order of students’ actions in learning system activity logs [21, 29, 33]. Examples of sequential pattern extraction and modelling have examined the temporality of students’ actions in order to gain insights into: the development of strategies, or to differentiate or group students who show similar behaviours [21]; student’s strategies [e.g., 31], students profiles [e.g., 34], conversation patterns [e.g., 30], the temporal evolution of student’s strategies [e.g., 22], or to compare cohorts of students by identifying the actions that differentiate them according to their expertise [19]. To a lesser extent, sequence mining techniques have also been used to focus on understanding the evolution of the artefacts created by the students [e.g., 29, 34].

The potential of these techniques for learning analytics is that analysing sequences of events can be a quite generic approach to consider the temporality for distilling interesting patterns where the sequence of occurrence (absence or dispersion) of certain event can be crucial for learning. We propose that such techniques could have potential to provide insights into student writing, and in the light of the above literature analysis, to provide feedback on
the rhetorical structure of a text. Although, similar sequence or process mining techniques have, to our knowledge, not yet been applied to student writing, some techniques applied to event data could easily be operationalised to uncover patterns in texts. To summarise, the preliminary work reported here motivates a theoretically sound rationale for the application of sequence mining to student writing, which we hope other researchers can use as a reference, and establish the steps needed to prepare data to maximise the opportunities of finding useful insights.

4. APPROACH

This section describes our current work aimed at analysing sequences and dispersion of moves in student writing. Below, we illustrate our approaches for preparing (pre-processing) and analysing (mining) the dataset based on our specific case of rhetorical moves in student writing. However, our claim is that the general approach described is applicable to other contexts, and is agnostic regarding the kind of analytic approach taken to identifying ‘moves’ in a text. We finalise the section with an illustrative example of cohort analysis of rhetorical moves dispersion.

4.1 Dataset: Annotation of Rhetorical Moves

Analysis is underway on a dataset from multiple disciplinary genres (law, accounting, and biology), annotated using the Xerox Incremental Parser’s (XIP) instantiation in tools for feedback on analytical academic writing [23, 39, 40]. Within the ‘Academic Writing Analytics’ (AWA) tool developed at UTS, the analytic parser is designed to detect rhetorical ‘steps’, inspired by Swales’ moves, that indicate specific rhetorical functions falling under the general moves described above. These steps are then – through AWA – highlighted within a submitted document, in order to give students feedback on the rhetorical structure of their text.

The analytical module labels thus include: Summarising issues (describing the article’s plan, goals, and conclusions) (S), describing Background knowledge (B), Contrastting ideas (C), Emphasising important ideas (E), mentioning Novel ideas (N), pointing out Surprising facts, results, etc. (P), describing an open Question or insufficient knowledge (Q), and recognising research Trends (T). Thus, as in Figure 1, each sentence in a document is labelled with none, or with one or more of these ‘steps’.

[Figure 1 – An Example AWA Report]

4.2 Dataset Preparation Approaches

Texts submitted to AWA are processed, with the marked up documents displayed to students in a ‘report’ which highlights individual sentences that display indicators of particular rhetorical moves. In order to process the student texts the original files (docs or pdf) were cleaned to remove: student IDs and names; headings; preface (such as cover sheets, or the essay prompt); end-matter (such as reference lists), and figures, tables, and captions. In addition, lists were converted to paragraph text. These texts were analysed using the AWA tool, with outputs in JSON format thus, documents submitted to AWA can be represented as sets of ordered sentences, with each sentence annotated with particular features – including which (if any) rhetorical moves it appears to exemplify. These texts may then be represented using a few different approaches, with important implications for the unit of analysis – for example, whether to treat the whole-text as a single unit, or to analyse meaningful sub-sections (such as the paragraph, or headed-section).

Essays can thus be represented as ordered lists of sentence-types. We can consider all the sentences or rather only those sentences for which ‘moves’ are identified (i.e., ignoring ‘blank’ sentences) – levels 1 and 2 in Figure 2. In addition, the repetition of particular features (including ‘blank’ sentences) might be collapsed into a single item, or treated separately. In addition, representations may capture elements of the process of writing a text – the key sections or elements that a text is built up from, for example, the paragraph, or introduction/body/conclusion – as in the right most two representations in Figure 3. In this latter model, sentences then, might either be treated as ordered or unordered elements within the constituent parts. That is, they can be treated as components of the sentences seen in the left most columns, or treated such that the order of the items within the sub-section does not matter.

4.3 Analysis (Mining) Approaches

Given the relationship between the XIP identified rhetorical moves within the AWA tool, and Swales’ rhetorical moves, our hypothesis is that methods to treat the sequences of rhetorical moves temporally may yield insight and aid us in designing actionable feedback for learners. Specifically, we are orienting our analysis around four approaches to the data analysis (which – as described below – may not be mutually exclusive):

Analysis of Sub-sequences: commonly occurring throughout a text – for example, the recurrence of “background”, “contrast”; “novel” moves in sequences indicating a typical analysis style sequence. Analysis based on this approach – for example using TraMineR’s analysis of frequent sub-sequences [16] – could indicate patterns in the ways in which groups of texts use particular structures in their text as indicated by their patterns of rhetorical move. This analysis could be conducted on any of the levels of data segmentation in Figure 2.

Analysis of Dispersion: or weighting of moves across sub-sections of a text (e.g. introduction, body, conclusion) – for example, a tendency for ‘background’ moves to appear more in the introduction, with ‘emphasising’ moves more in a conclusion. This type of analysis could also indicate the dispersion of sub-sequences throughout a text. Analysis based on this approach could indicate the ways that texts are structured to show how
sections of text fulfill particular functions. Analysis of this sort could be conducted on levels 3 & 4 of representing the data, as it requires a way for weighting the moves within particular sub- sections.

Analysis of Associations between moves that co-occur in meaningful sub-sections of text (e.g. the paragraph level), in non-ordered ways. For example, a tendency for ‘novel’ and ‘surprise’ moves to co-occur. This type of analysis could indicate that particular sub-sections of text have characteristic moves – but that these moves can occur in a variety of orders. Analysis of this sort could be conducted on any of the levels of representing the data.

Analysis of Processes of moves occurring over a whole text – for example, a text might include a cycle of sections weighted towards a particular move (as in the ‘dispersion’ analysis), in sequences – for example, more ‘background’ moves, followed by more ‘question’ moves, followed by more ‘novel’, although within each element other moves might occur. For example, we might see a repeated pattern of shifts from mostly ‘background’ steps to mostly ‘question’ steps, to mostly ‘trends’ steps, as in discussion papers that point to emerging research directions based on questions in the earlier literature. Analysis of this sort could be conducted on any (1–4) of the levels of representing the data.

4.4 Illustrative Example: Cohort Analysis of Rhetorical Moves Dispersion

Here we present a preliminary example of the potential of analysing the dispersion of rhetorical moves in student’s texts. Figure 443 visualises the difference in rhetorical moves for student’s texts of two cohorts: those that received high distinction (HD) and pass (P) marks. For illustrative purpose, this simple visualisation shows all the rhetorical moves of the texts divided at a section level, with each student’s text arbitrarily divided in quantiles to explore the dispersion of the rhetorical moves.

The HD texts present a larger mix of rhetorical moves, particularly in the first and last quintiles (see particularly Q1 and Q2 for texts B and C). Notably, they all commenced with an EMPHASIS move followed by a combination of SUMMARY. CONTRAST sentences and other moves. In Q4 and Q5 these texts also showed a quite varied combination of moves. By contrast, the P texts, overall, showed fewer, and more dispersed, rhetorical moves. Notably, the three P examples present some moves in the middle of the text (at Q3) which are not observed in the HD texts. Additionally, these texts include a smaller range of moves either at the beginning (Q1 for F) or at the end (Q5 in D and E) of the text. This illustrates the potential insights that can be gained by analysing sequencing and dispersion of rhetorical moves. Finding patterns of this type in larger cohorts can be useful to generate the means for providing automated or hybrid feedback to the students about their writing.

This paper has sought to present an account of sequential and process analyses of writing, using data from our own research to highlight the ways in which student-texts might be treated using these approaches. In doing so, we foreground the potential of different types of analysis and data representation. There is potential to extend this work to feedback to students and instructors regarding the structure of their written work. Analyses are ongoing to demonstrate this potential empirically.

[6] Cortes, V. 2013. The purpose of this study is to: Connecting lexical bundles and moves in research article introductions. Journal of English for academic purposes. 12, 1 (2013), 33–43.


