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
{Simon.Knight, Roberto.Martinez-Maldonado, Andrew.Gibson, Simon.BuckinghamShum}@uts.edu.au

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 undercover 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: 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).

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 by our developing work on a corpus annotated with 'rhetorical moves' and their analysis (§4). 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 (§4.2-4.3).

2. SEQUENCE AND PROCESS ANALYSES OF STUDENT WRITING
A small body of work has explored 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 revisioning 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. premise, argumentative move, claim) are fundamental to definitions of "argument" and the computational modelling of argumentation [e.g., 15, 45], thus lending itself to sequence analysis for identification of argument structures and arguments in use through the analysis of argument constituent parts [e.g., 28].

Obviously the 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]. Linguistics 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 written products, a limited body of work has investigated the sequential structure of 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 infrequently 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

That work [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. Extant 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, extant 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 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 undercover 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 presents work 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), *Contrasting* 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'.
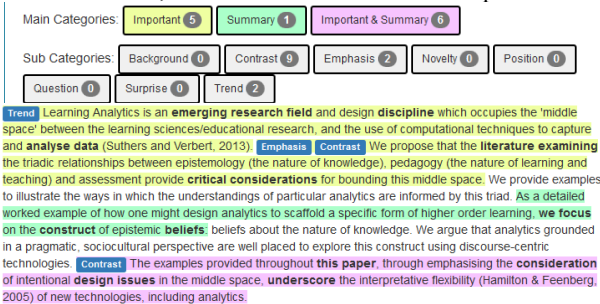

Figure 1 – An Example AWA Report

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 (docx 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 2 (levels 3 and 4). 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 sequences seen in the left most columns, or treated such that the order of the items within the sub-section does not matter.
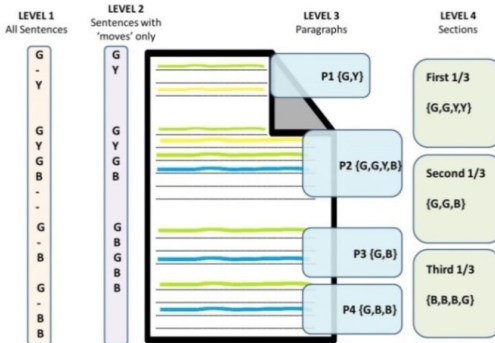

Figure 2 – Levels for Representing/Aggregating Student's Writing: 1) Sentences (all), 2) Sentences marked as 'moves', 3) Paragraphs, and 4) Sections

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 'emphasizing' 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 fulfil 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 3 shows the rhetorical moves for 6 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 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 may be useful to generate the means for providing automated or hybrid feedback to the students about their writing.

5. CONCLUSIONS

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 draw on these analyses to feedback to students and instructors regarding the structure of their written work. Analyses are ongoing to demonstrate this potential empirically.

6. REFERENCES

[1] Béchet, N., Cellier, P., Charnois, T. and Crémilleux, B. 2012. Discovering linguistic patterns using sequence mining. *International Conference on Intelligent Text Processing and Computational Linguistics* (2012), 154–165.
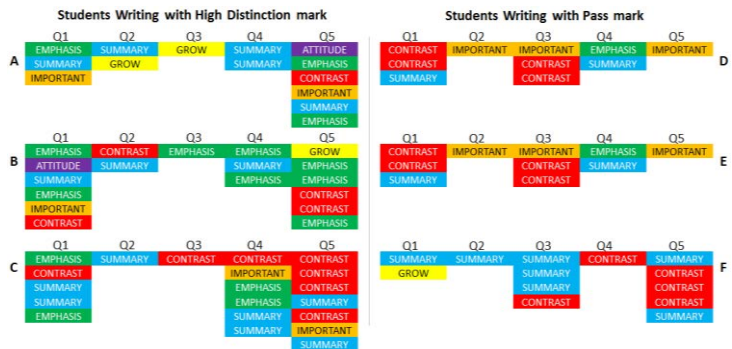
Figure 4 – Visualisations of 6 pieces of Science students' writing that represent the potential of cohort analysis of rhetorical moves dispersion

[2] Bruce, B.C. 1972. A model for temporal references and its application in a question answering program. *Artificial intelligence*. 3, (1972), 1–25.

[3] Buckingham Shum, S., Knight, S., McNamara, D., Allen, L., K.., Betik, D. and Crossley, S. 2016. Critical Perspectives on Writing Analytics. (Edinburgh, UK, 2016), 481–483.

[4] Calvo, R.A., O'Rourke, S.T., Jones, J., Yacef, K. and Reimann, P. 2011. Collaborative writing support tools on the cloud. *IEEE Transactions on Learning Technologies*. 4, 1 (2011), 88–97.

[5] Chen, B., Wise, A.F., Knight, S. and Cheng, L., K.. 2016. It's About Time: Putting Temporal Analytics into Practice: The 5th International Workshop on Temporality in Learning Data. (Edinburgh, UK, 2016), 488–489.

[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.

[7] Cotos, E. 2009. Designing an intelligent discourse evaluation tool: Theoretical, empirical, and technological considerations. Iowa State University.

[8] Cotos, E. and Huffman, S. 2013. Learner fit in scaling up automated writing evaluation. *International Journal of Computer-Assisted Language Learning and Teaching (IJCALLT)*. 3, 3 (2013), 77–98.

[9] Cotos, E. and Pendar, N. 2016. Discourse classification into rhetorical functions for AWE feedback. *Calico Journal*. 33, 1 (2016), 92.

[10] Crossley, S., Dascalu, M., Trausan-Matu, S., Allen, L. and McNamara, D. 2016. Document Cohesion Flow: Striving towards Coherence. *Cognitive Science Society* (2016).

[11] Crossley, S., Dempsey, K. and McNamara, D. 2011. Classifying paragraph types using linguistic features: Is paragraph positioning important? *Journal of Writing Research*. 3, 2 (Dec. 2011), 119–143.

[12] Crossley, S.A., Defore, C., Kyle, K., Dai, J. and McNamara, D.S. 2013. Paragraph Specific N-Gram Approaches to Automatically Assessing Essay Quality. *EDM* (2013), 216–219.

[13] Dayrell, C., Candido Jr, A., Lima, G., Machado Jr, D., Copestake, A.A., Feltrim, V.D., Tagnin, S.E. and Aluísio, S.M. 2012. Rhetorical Move Detection in English Abstracts: Multi-label Sentence Classifiers and their Annotated Corpora. *International Conference on Language Resources and Evaluation* (Istanbul, Turkey, 2012).

[14] Denicia-Carral, C., Montes-y-Gómez, M., Villaseñor-Pineda, L. and Hernández, R.G. 2006. A text mining approach for definition question answering. *Advances in Natural Language Processing*. Springer. 76–86.

[15] Feng, V.W. and Hirst, G. 2011. Classifying arguments by scheme. *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies-Volume 1* (2011), 987–996.

[16] Gabadinho, A., Ritschard, G., Mueller, N.S. and Studer, M. 2011. Analyzing and visualizing state sequences in R with TraMineR. *Journal of Statistical Software*. 40, 4 (2011), 1–37.

[17] Guo, Y., Korhonen, A. and Poibeau, T. 2011. A weakly-supervised approach to argumentative zoning of scientific documents. *Proceedings of the Conference on Empirical Methods in Natural Language Processing* (2011), 273–283.

[18] Hastings, P., Hughes, S., Britt, A., Blaum, D. and Wallace, P. 2014. Toward Automatic Inference of Causal Structure in Student Essays. *Intelligent Tutoring Systems*. S. Trausan-Matu, K.E. Boyer, M. Crosby, and K. Panourgia, eds. Springer International Publishing. 266–271.

[19] Jiang, Y., Paquette, L., Baker, R.S. and Clarke-Midura, J. 2015. Comparing Novice and Experienced Students within Virtual Performance Assessments. *International Educational Data Mining Society*. (2015).

[20] Kale, M.S., Palshikar, G.K., Chhajed, S. and Deshpande, L. 2005. Data Mining over Textual Data. *TACTiCS – TCS Technical Architects' Conference* (London, UK, 2005).

[21] Kinnebrew, J. and Biswas, G. 2012. Identifying learning behaviors by contextualizing differential sequence mining with action features and performance evolution. *Educational Data Mining 2012* (2012).

[22] Kinnebrew, J.S., Segedy, J.R. and Biswas, G. 2014. Analyzing the temporal evolution of students' behaviors in open-ended learning environments. *Metacognition and Learning*. 9, 2 (2014), 187–215.

[23] Knight, S., Buckingham Shum, S., Ryan, P., Sándor, Á. and Wang, X. Forthcoming. Academic Writing Analytics for Civil Law: Participatory Design Through Academic and Student Engagement. *International Journal of Artificial Intelligence in Education*. (Forthcoming).

[24] Knight, S. and Littleton, K. 2015. Discourse-Centric Learning Analytics: Mapping the Terrain. *Journal of Learning Analytics*. 2, 1 (2015), 185–209.

[25] Knight, S., Wise, A.F., Chen, B. and Cheng, B.H. 2015. It's About Time: 4th International Workshop on Temporal Analyses of Learning Data. (Poughkeepsie, NY, USA, Mar. 2015), 388–389.

[26] Lawrence, J. and Reed, C. 2015. Combining argument mining techniques. *NAACL HLT 2015*. (2015), 127.

[27] Liu, M., Calvo, R.A. and Pardo, A. 2013. Tracer: A Tool to Measure and Visualize Student Engagement in Writing Activities. *2013 IEEE 13th International Conference on Advanced Learning Technologies* (Jul. 2013), 421–425.

[28] Madnani, N., Heilman, M., Tetreault, J. and Chodorow, M. 2012. Identifying high-level organizational elements in argumentative discourse. *Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies* (2012), 20–28.

[29] Maldonado, R.M., Yacef, K., Kay, J., Kharrufa, A. and Al-Qaraghuli, A. 2010. Analysing frequent sequential patterns of collaborative learning activity around an interactive tabletop. *Educational Data Mining 2011* (2010).

[30] Martinez-Maldonado, R., Dimitriadis, Y., Martinez-Monés, A., Kay, J. and Yacef, K. 2013. Capturing and analyzing verbal and physical collaborative learning interactions at an enriched interactive tabletop. *International Journal of Computer-Supported Collaborative Learning*. 8, 4 (2013), 455–485.

[31] Martinez-Maldonado, R., Yacef, K. and Kay, J. 2013. Data Mining in the Classroom: Discovering Groups' Strategies at a Multi-tabletop Environment. *Educational Data Mining 2013* (2013).

[32] Myers, J.C., McCarthy, P.M., Duran, N.D. and McNamara, D.S. 2011. The bit in the middle and why it's important: a computational analysis of the linguistic features of body paragraphs. *Behavior research methods*. 43, 1 (2011), 201–209.

[33] Pechenizkiy, M., Trcka, N., Vasilyeva, E., van Aalst, W. and De Bra, P. 2009. Process mining online assessment data. *Educational Data Mining* (2009), 279–288.

[34] Perera, D., Kay, J., Koprinska, I., Yacef, K. and Zaïane, O.R. 2009. Clustering and sequential pattern mining of online collaborative learning data. *Knowledge and Data Engineering, IEEE Transactions on*. 21, 6 (2009), 759–772.

[35] Plantevit, M., Charnois, T., Klema, J., Rigotti, C. and Crémilleux, B. 2009. Combining sequence and itemset mining to discover named entities in biomedical texts: a new type of pattern. *International Journal of Data Mining, Modelling and Management*. 1, 2 (2009), 119–148.

[36] Roscoe, R., Crossley, S., Weston, J. and McNamara, D. 2011. Automated Assessment of Paragraph Quality: Introduction, Body, and Conclusion Paragraphs. *Twenty-Fourth International FLAIRS Conference* (Mar. 2011).

[37] Roscoe, R.D., Snow, E.L., Allen, L.K. and McNamara, D.S. 2015. Automated detection of essay revising patterns: applications for intelligent feedback in a writing tutor. *Cognition and Learning*. 10, 1 (2015), 59–79.

[38] Salager-Meyer, F. 1992. A text-type and move analysis study of verb tense and modality distribution in medical English abstracts. *English for specific purposes*. 11, 2 (1992), 93–113.

[39] Simsek, D., Buckingham Shum, S., Sandor, A., De Liddo, A. and Ferguson, R. 2013. XIP Dashboard: visual analytics from automated rhetorical parsing of scientific metadiscourse. (Leuven, Belgium, 2013).

[40] Simsek, D., Sandor, A., Shum, S.B., Ferguson, R., De Liddo, A. and Whitelock, D. 2015. Correlations between automated rhetorical analysis and tutors' grades on student essays. *Proceedings of the Fifth International Conference on Learning Analytics And Knowledge* (2015), 355–359.

[41] Southavilay, V., Yacef, K., Reimann, P. and Calvo, R.A. 2013. Analysis of Collaborative Writing Processes Using Revision Maps and Probabilistic Topic Models. *Proceedings of the Third International Conference on Learning Analytics and Knowledge* (New York, NY, USA, 2013), 38–47.

[42] Sun, G., Liu, X., Cong, G., Zhou, M., Xiong, Z., Lee, J. and Lin, C.-Y. 2007. Detecting erroneous sentences using automatically mined sequential patterns. *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics* (Prague, Czech Republic, 2007), 81–88.

[43] Swales, J.M. 1990. *Genre analysis: English in academic and research settings*. Cambridge University Press.

[44] Teufel, S. and Moens, M. 2002. Summarizing scientific articles: experiments with relevance and rhetorical status. *Computational linguistics*. 28, 4 (2002), 409–445.

[45] Walton, D. 2012. Using argumentation schemes for argument extraction: A bottom-up method. *International Journal of Cognitive Informatics and Natural Intelligence (IJCINI)*. 6, 3 (2012), 33–61.