(2017). Towards reflective writing analytics: Rationale, methodology, and preliminary results. *Journal of Learning Analytics*, *4*(1), 58–84. http://dx.doi.org/10.18608/jla.2017.41.5

Towards Reflective Writing Analytics: Rationale, Methodology and Preliminary Results

Simon Buckingham Shum
Connected Intelligence Centre, University of Technology Sydney, Australia

Ágnes Sándor
Xerox Research Centre Europe, France

Rosalie Goldsmith
Institute for Interactive Media in Learning, University of Technology Sydney, Australia

Randall Bass, Mindy McWilliams
Georgetown University, USA

Simon.BuckinghamShum@uts.edu.au

ABSTRACT: When used effectively, reflective writing tasks can deepen learners' understanding of key concepts, help them critically appraise their developing professional identity, and build qualities for lifelong learning. As such, reflective writing is attracting substantial interest from universities concerned with experiential learning, reflective practice, and developing a holistic conception of the learner. However, reflective writing is for many students a novel genre to compose in, and tutors may be inexperienced in its assessment. While these conditions set a challenging context for automated solutions, natural language processing may also help address the challenge of providing real time, formative feedback on draft writing. This paper reports progress in designing a writing analytics application, detailing the methodology by which informally expressed rubrics are modelled as formal rhetorical patterns, a capability delivered by a novel web application. Preliminary tests on an independently human-annotated corpus are encouraging, showing improvements from the first to second version, but with much scope for improvement. We discuss a range of issues: the prevalence of false positives in the tests, areas for future technical improvements, the issue of gaming the system, and the participatory design process that has enabled work across disciplinary boundaries to develop the prototype to its current state.

Keywords: Learning analytics, education, writing analytics, reflection, natural language processing, reflective move, rhetoric

1 ACADEMIC REFLECTIVE WRITING

Reflection has long been regarded as a key element in student learning and professional practice in higher education (Boud, Keogh, & Walker, 1985; Hatton & Smith, 1995; Rodgers, 2002a; Ryan, 2011). It

(2017). Towards reflective writing analytics: Rationale, methodology, and preliminary results. *Journal of Learning Analytics*, *4*(1), 58–84. http://dx.doi.org/10.18608/jla.2017.41.5

can allow students a window into their developing professional identity (Reidsema, Goldsmith, & Mort, 2010), deepen understanding of key concepts (Scouller, 1998), and provide opportunities for lifelong learning (Ryan, 2011). However, it has been so broadly interpreted and implemented in the university curriculum that the concept of reflection has become attenuated (Webster-Wright, 2013). Because of such broad interpretations, clarifying what is meant by reflection is no easy task (Rodgers, 2002a), but is critical to meaningful discussion. The definition by Boud et al. (1985) provides a useful perspective:

> Reflection is an important human activity in which people recapture their experience, think about it, mull over and evaluate it. It is this working with experience that is important in learning. (p. 43)

Reflection is thus regarded as an intrinsic element of learning, especially of experiential learning in professional degree programs such as teacher education, nursing, engineering, and architecture. As reflection is a social cognitive process, one of the challenges when using it as a tool for learning is to find ways in which students can demonstrate their reflective activities (Boud et al., 1985; Hatton & Smith, 1995). Reflective writing tasks are the most common form of implementing reflective activities in the university curriculum, as writing is still the main form of assessment in higher education, notwithstanding a number of debates surrounding the practice of reflective writing. These debates include issues such as how such tasks should be incorporated into the curriculum, how such writing should be taught or developed, and how — or indeed whether — reflective writing should be assessed (Boud et al., 1985; Sumsion & Fleet, 1996). Extended writing is of course just one window into the mind of a student's ability to reflect; reflective video annotation with brief notes about one's own or a peer's performance is another such window (Hulsmann & van der Floodt, 2015).

This paper is organized as follows. In the next section, we recognize that reflective writing is for many students, and educators, a novel genre to compose in, and to assess. We then introduce the particular contexts in which we are using reflective writing (Section 3). This sets the background for considering the potential for automated writing analytics to provide timely, personalized, formative feedback to students on their drafts. The technical platform we are developing to analyze reflective writing is presented (Section 4), before describing the methodology by which we move from rubrics, to rhetorical patterns (Section 5) that are implemented in a parser. Two iterations of the parser development are then detailed (Section 6). The discussion reflects on the complexities of classifying reflective statements, and the importance of a participatory process for establishing trust among the diverse stakeholders in an analytics ecosystem (Section 7).

2 ASSESSMENT CHALLENGES

As noted above, reflective writing is a novel genre for many students; thus learning to write reflectively at university can present a number of challenges. One challenge is to understand the purpose of reflection and reflective writing; and to take it as "seriously" as more traditionally "academic" genres

(2017). Towards reflective writing analytics: Rationale, methodology, and preliminary results. *Journal of Learning Analytics*, *4*(1), 58–84. http://dx.doi.org/10.18608/jla.2017.41.5

such as reports, case studies, and essays. Educators do not always make the purpose explicit, and perhaps they themselves are not always clear about what they are asking their students to do (Rodgers, 2002a). Another challenge for students is to decipher what is expected in a piece of reflective writing, both in content and in style. The requirement for a personal tone in an academic assignment can often cause confusion and uncertainty for student writers.

The affective domain of reflection, and of reflective writing, is a relatively unexplored area (but see Boud et al., 1985; Dewey 1933 in Rodgers 2002a), although most proponents of reflective writing acknowledge its more personal tone. The affective domain is most often the trigger for reflection: "I feel excited, nervous, uncomfortable, embarrassed, distressed, elated, curious, concerned — why do I feel this way? What is causing this disruption?" An emotional response to a learning event is frequently the first signal to the learner that something of significance has occurred, or that a shift of some kind needs to take place. Such responses and the type of writing that explores and discusses these responses is in strong contrast to the types of academic genres that students write in as part of their studies, and especially as part of their assessment.

The assessment of reflective writing is less straightforward than for more familiar forms of analytical academic writing. This is in part because reflective writing is different in nature and purpose; its intention is to communicate a subjective, personal, individual interpretation of experiences and what has been learned from them. Students are encouraged to consider their mistakes and demonstrate changes in points of view rather than present the correct answer. Another potentially problematic aspect of assessing reflective writing is the different perspectives (of academics and students) on what reflective writing could or should be. A shared understanding of what constitutes a deep or superficial reflection is critical to valid and reliable assessment, but the literature indicates that this has been an ongoing challenge. Inter-coder reliability has been particularly difficult to establish (Hatton & Smith, 1995; Sumsion & Fleet, 1996).

Related to this is the need for a shared language to teach and assess reflective writing, as identified by Ryan (2011) in a project specifically intended to develop the teaching of reflective writing across a number of disciplines in an Australian university (Ryan & Ryan, 2012). Many academics lack the meta-language to identify or explain what they regard as key elements of deep reflective writing. They are therefore unable either to give clear directions to students about how to approach a reflective writing task, or to justify the marks that they give to student assignments.

Boud and Walker (1998) put forward the argument that, since reflective writing is very different in nature and purpose from analytical academic writing, it should be assessed using criteria sensitive to that particular genre (p. 194). In their seminal paper on how and whether to assess reflective writing tasks, Sumsion and Fleet (1996) make the important point that some students may reflect deeply but not have mastery of the genre of reflective writing, whereas other students with stronger writing skills or abilities to write reflectively may appear to be reflecting without actually doing so (p. 124). This is an

(2017). Towards reflective writing analytics: Rationale, methodology, and preliminary results. *Journal of Learning Analytics*, *4*(1), 58–84. http://dx.doi.org/10.18608/jla.2017.41.5

aspect of reflective writing that is difficult to resolve, but is one that is worth trying to parse in analysis. Additionally, reflective writing often asks students to reflect on experiences in a personal way. Therefore, they must decide to what degree they wish to disclose their uncertainties and vulnerabilities, and then express that appropriately in academic reflective writing; this will be assessed as a strength rather than a weakness.

Thus it can be seen that although reflective writing can be a powerful tool in student learning in the higher education context, its practice and assessment are by no means clear-cut. On the one hand, there is a risk that students have not been properly introduced to it as a new form of writing that is relevant to their studies, and will approach reflective writing tasks in a strategic or perfunctory manner as "simply another assignment to complete as efficiently as possible." The evidence in the literature cited above is that students typically respond with superficial descriptions of their experiences or with broad statements such as "I learned a lot." On the other hand, as detailed below, when we consider assessment, it is not straightforward to establish a shared understanding amongst academics (not to mention students) of what appropriate reflection is when expressed in academic writing, and how it can be developed and assessed.

Lastly, a critical challenge to address is that of capacity to provide rigorous assessment and personalized feedback at scale (cf. the contexts at the University of Technology Sydney and Georgetown University in Washington, DC, introduced next). When teaching a large course, the assessment of any written assignment or paper becomes a daunting task, now made more complex by the unfamiliar genre of academic reflective writing. If instructors do not know how to provide appropriate feedback and grading, this risks confirming in students' minds that this novel kind of reflection is peripheral to, or an interesting diversion from, the "real learning" that they signed up for.

In light of evidence about the benefits of reflective writing reviewed initially, these complexities do not dissuade many educators from using reflective writing as a way to help students engage in deeper internalization and meaning-making of their experiences, as interpreted and analyzed through the lens of theory or discipline. However, our goal is to see how we may lower these "entry barriers" to shifting assessment towards deeper reflection on authentic learning.

This sets the challenging context into which we now introduce learning analytics. Our working hypothesis is that writing analytics in principle could be an enabler if a tool can help educators adopt new practices with reflective writing, with enhanced formative feedback available to students to help build their ability. Is reflective writing, in all its complexity, amenable to natural language processing (NLP), to deliver meaningful feedback?

(2017). Towards reflective writing analytics: Rationale, methodology, and preliminary results. *Journal of Learning Analytics*, *4*(1), 58–84. http://dx.doi.org/10.18608/jla.2017.41.5

3 REFLECTIVE WRITING CONTEXTS

3.1 Reflective Writing for Engineers (UTS)

At the University of Technology Sydney, all engineering students in the 4-year degree program undertake two 6-month internships that are part of the practice program. At the completion of each internship, students are required to submit a reflective report that details changes in their professional, personal, and technical awareness. The cohort size is approximately 200 per semester; the reports are expected to be 40–50 pages and hence are very time-consuming to mark. It is difficult for tutors to provide formative feedback on drafts during the semester, both because of the size of the cohort and because the subject is delivered in block mode, where students attend intensive all-day sessions over the semester rather than weekly classes. An initiative is now under way to develop finer-grained assessment and grading of reflective writing, which contributes to the context for the writing analytics work reported here.

3.2 Reflective Writing for "Formation" (GU)

For about two years, the *Formation by Design* project[1] at Georgetown University (GU) has been working (in collaboration with others, including UTS), to consider how the concept of "formation" should shape the university experience — specifically, how do we define, intentionally design for, and assess the qualities of a fully formed person? As the project defines it: "The concept of formation is at the heart of an education dedicated to shaping students to be fully human, to cultivating their authentic selves, and to inhabiting a sense of personal responsibility for improving the world." The importance of redefining metrics and analytics sits at the heart of the work: "Learning — and especially 'learner-centered' — analytics hold much promise as a mechanism for integrating qualitative and quantitative measures of formation, as well as visualizing and feeding meaningful data back to stakeholder groups at every level of the educational ecosystem."

A key approach in this work is the process of internal reflection that integrates new knowledge and experiences, and creates meaning from these. Reflective writing, used in academic settings such as course work following experiential learning, is a commonly used technique to both provoke the action of reflection, and capture the product of reflection for interpretation by another person, most often the course instructor, who uses this product to interpret and assess the learning and change that has taken place.

The Engelhard Project for Connecting Life and Learning at GU, which aims to increase student well-being and deepen academic engagement, has been using reflective writing for ten years in over 325 courses, resulting in a corpus of thousands of student reflections in over 28 disciplines. A sample of this corpus,

[1] https://futures.georgetown.edu/formation

(2017). Towards reflective writing analytics: Rationale, methodology, and preliminary results. *Journal of Learning Analytics*, *4*(1), 58–84. http://dx.doi.org/10.18608/jla.2017.41.5

taken from courses in Biology, Health Studies, Philosophy, Psychology, and Sociology, was used in the collaborative effort described in this paper to explore an analytics-supported approach to assessing learning, growth, and change in student reflective essays.

4 MODELLING REFLECTIVE WRITING

4.1 NLP Platform: XIP

We use the Xerox Incremental Parser (XIP) for automated reflective writing analysis (Aït-Mokhtar, Chanod, & Roux, 2002).[2] XIP is a linguistic analysis engine (including some statistical but primarily rule-based components) and it contains a rule writing formalism, which has provided the platform for the development of high-performance, deep dependency parsing of general English texts. The input to the analysis is free text, which is incrementally processed by consecutive NLP steps: from segmentation into sentences and lexical units, through part-of-speech disambiguation, to extracting syntactic dependency relationships among the lexical elements. Besides syntactic analysis, XIP processing performs general semantic analysis functions such as named entity recognition (Brun & Hagège, 2004) and semantic normalization (Brun & Hagège, 2003). The maturity of the syntactic and semantic parsing capability is evidenced by its applications for a wide variety of NLP tasks including information extraction (Huang, ten Teije, van Harmelen, & Aït-Mokhtar, 2014), sentiment analysis (Brun, Popa, & Roux, 2014), and discourse analysis (Sándor, 2007).

XIP includes a "salient sentences" module that models and detects sentences conveying relevant rhetorical moves in *analytical* writing genres like scientific and scholarly research articles, and research reports (De Liddo, Sándor, & Buckingham Shum, 2012; Lisacek, Chichester, Kaplan, & Sándor, 2005; Sándor & Vorndran, 2009). It is based on reliable dependency parsing, and an integrated set of NLP capabilities that provide the necessary resources to build patterns for capturing rhetorical moves of analytical writing. Simsek, Buckingham Shum, Sándor, De Liddo, and Ferguson (2013) provide a more detailed rationale for the use of the analytical writing parser in education, and the description of a prototype dashboard. Simsek et al. (2015) report a preliminary evaluation in the context of an English literature student assignment, and Knight, Buckingham Shum, Ryan, Sándor, and Wang (in press) report a more detailed evaluation with Civil Law students. The reflective writing parser documented in this paper is an extension of this XIP module.

4.2 AWA: A Writing Analytics Application Using XIP

This work is part of a broader effort at UTS to rapidly prototype writing analytics of different types together with staff and students. A collaborative effort between learning analytics specialists and academic staff resulted in the creation of a web application called *Academic Writing Analytics (AWA)* as

[2] *Xerox Incremental Parser*. Open Xerox Documentation: https://open.xerox.com/Services/XIPParser

(2017). Towards reflective writing analytics: Rationale, methodology, and preliminary results. *Journal of Learning Analytics*, *4*(1), 58–84. http://dx.doi.org/10.18608/jla.2017.41.5

an educational application using XIP's services.[3] This enabled a piece of writing to be submitted for analysis, and the raw output from the parser was rendered in AWA as interactive highlighted text (illustrated in Figure 1).

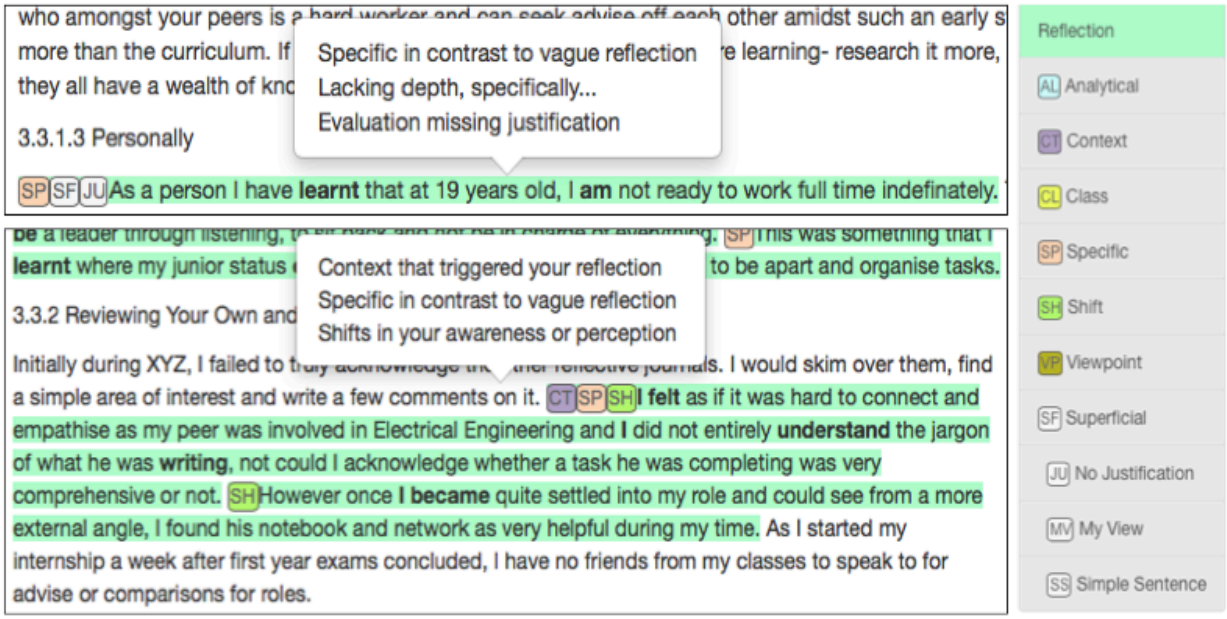

Figure 1: At the time of this research, AWA's user interface highlighted sentences in the student's text matching XIP's rhetorical patterns as shown. "Function Keys" such as SH (SHIFT) signalled the function that the sentence had been classified as playing, and mousing over the highlight displayed a prompt with the meaning of the F key.

4.3 Related Approaches

Although reflective writing has been studied widely, little work has been devoted to its automated analysis. Owing to the complexity of describing or formalizing the features of reflective writing, the constitution of annotated corpora and establishing evaluation measures are major challenges for the task. Another specific feature of reflective writing pieces is that unlike analytical texts, the overall structure is not standardized. Consequently, text structure is not leveraged in the analysis. We are at present aware of only two other learning analytics projects related to reflective writing, proposing different methods for reflection detection, corpus constitution, and evaluation.

Ullmann, Wild, and Scott (2012) developed a rule-based categoriser that decides if a text is reflective or not. Based on theoretical research in reflective writing, the authors proposed a list of five "elements of reflection": *Description of experience, Personal experience, Critical analysis, Taking perspectives into*

[3] *Academic Writing Analytics* information: https://utscic.edu.au/tools/awa

(2017). Towards reflective writing analytics: Rationale, methodology, and preliminary results. *Journal of Learning Analytics*, *4*(1), 58–84. http://dx.doi.org/10.18608/jla.2017.41.5

account, and *Outcome of reflective writing*. These elements are associated with sets of indicators used in 16 rules to detect reflective sentences, such as a rule for detecting *Description of an experience* is "Past tense sentence with self-related pronoun as subject." Eight different resources and tools serve as dedicated annotators that provide input for the rules, such as the Stanford Parser to perform syntactic parsing for identifying subjects of sentences, and a self-reference annotator that contributes with a list of lexical elements conveying self-reference.

The whole system is integrated within the UIMA framework.[4] The input texts are categorized as either reflective or not reflective according to the presence and the quantity of the detected elements of reflection. The system parameters were developed based on 10 prototypical reflective texts, and the test was carried out by crowdsourcing with paid annotators (via Amazon Mechanical Turk) who evaluated the presence of the reflective elements in texts. The evaluation texts are a collection of blog posts, and their topic is not specified in the paper. The results showed a positive correlation between the reflective features identified by the annotators, and the texts categorized as reflective by the parser.

Ullmann et al.'s rule-based methodology is similar to ours, and the elements of reflection that they identify overlap with the rubrics and patterns described in Section 5.2. The major differences between the two systems are the modelling and the implementation frameworks as well as the evaluation method. Whereas Ullmann et al. build their system on independent indicators of reflective writing, we propose modelling reflective moves as rhetorical patterns. For the implementation, they use an array of tools for detecting the different indicators of the reflective elements, and an independent rule formalism, while XIP is a single, modular system implementing syntactic analysis, lexical resources, and the dependency rules that detect the reflective patterns. We cannot directly compare the performance results of the two parsers since the results reported in Ullmann et al. refer to a whole-document categorization task, while the task XIP performs is to detect and label reflective sentences without evaluating the whole document as reflective or not.

In contrast to Ullmann et al. (2012) and this paper, Gibson and Kitto (2015) do not focus on the fully automated detection of the linguistic indicators of academic reflective writing; instead, they aim to develop a way to model how NLP could support (not automate) the human identification of "anomalies" in a text, a potential ingredient in reflective writing: "Essentially, our objective was to outline the necessary steps that, given an anomaly in one context, allow a new context to be created in which that anomaly is resolved, without modifying the original context." Anomalies include student irony, sarcasm, and humour (e.g., *"I'm spending my weekend marking assignments. I love it — can't imagine doing anything else"*). Their *Anomaly Recontextualization* approach thus seeks to formalize the distinctive human ability to recognize and make sense of information that is apparently anomalous, until one reframes the context. Their underlying motivation is capturing subjective and affective features, which

[4] Apache Unstructured Information Management Architecture: https://uima.apache.org

(2017). Towards reflective writing analytics: Rationale, methodology, and preliminary results. *Journal of Learning Analytics*, *4*(1), 58–84. http://dx.doi.org/10.18608/jla.2017.41.5

are distinctive in reflective texts compared to analytical writing. They report preliminary results showing that when supervised, the model is capable of identifying different kinds of anomalies in student feedback, in relation to a student-supplied rating of "progress satisfaction," and an analyst supplied coding of "self–others balance."

Building on and extending this work, Gibson, Kitto, and Bruza (2016) have more recently proposed a conceptual model to explain the relationships between *reflection* and *metacognition* in learning. This motivates exploratory computational analysis of undergraduate reflective writing, in which a range of textual features is mapped to the model's constructs. Reflective writing, therefore, was analyzed in order to answer research questions about the validity of the conceptual model. They argue that this provides a conceptual foundation for the development of formative feedback to students, and identify as future work the need to evaluate their parsers on an independently annotated corpus from their development corpus.

5 ITERATIVE DESIGN METHODOLOGY

We now describe a rapid prototyping methodology for formalizing rubrics into executable patterns in XIP. The availability of an independently annotated corpus at Georgetown University offered the chance to compare the output of our system with human highlighting of sentences conveying relevant reflective moves. In the discussion, we reflect on whether this model could generalize to other contexts.

5.1 Start with Informal Rubrics

Rubrics are common in education, as an instructional and grading guide for students and graders as to what "good" looks like, sometimes mapped to different grades. In Australia, as in many other countries, universities have *Academic Language & Learning* (ALL) teams (or variations on that name) who build the capacity of academics and students to deploy language more effectively in their teaching and learning. We have found that ALL experts play an important role in the writing analytics design process. The first step was for the ALL expert affiliated with the UTS engineering faculty (Goldsmith), to provide a set of examples of the kinds of constructions that are typical signifiers of a reflective move. In order to develop a greater shared understanding amongst the engineering tutors in the practice program of what reflective writing is, and how it could be developed and assessed, the ALL expert had consulted with one of the subject coordinators. Through a combination of prior scholarship in the field to contextualize research for practitioners (Moon, 2010; Ryan, 2010), direct analysis of engineering students' reflective reports, and discussion with the subject coordinator, the ALL expert designed the rubric to identify linguistic features and textual moves commonly associated with deep or significant reflections (Table 1).

(2017). Towards reflective writing analytics: Rationale, methodology, and preliminary results. *Journal of Learning Analytics*, *4*(1), 58–84. http://dx.doi.org/10.18608/jla.2017.41.5

Table 1: Rubrics for Reflective Writing

1. Describing the context of the event that triggers the reflection (*why, when, where, who, how much, what*): the more detail the better, as long as the event is non-trivial
2. Expressions about learning something specific, e.g., *I learned that* (i.e., not merely "I learned a lot")
3. Expressions of reflecting specifically, e.g., *On reflection I could see that*
4. Expressions of increased confidence or ability, e.g., *I am more confident, am now able, feel/am comfortable, can plan, can utilize, can develop a strategy*
5. Expressions of experimentation and ability, e.g., *I tried, I tested, I experimented, I developed my capability to, I was/am able to, I was/am unable to, I practised, I asked, I sought advice, I overcame, I couldn't overcome*
6. Verbs that show awareness or shifts in perception, e.g., *I began to understand, I could see, I could visualize, I could perceive, I became aware, I became, I grew, I realized, I recognized*
7. Reference to the past: time markers and use of past tense (e.g., *when I started; before my internship*); shift between habitual past tense (e.g., *I used to*) and the present or the recent past (e.g., *since then I have*)
8. Reference to the present and future in the context of reflecting on changed behaviour, expectations or beliefs, e.g., *since, now, when, as it turned/turns out, it became clear*
9. Expressions of the unexpected and of prior assumptions, e.g., *I thought, I believed, I expected, I assumed, I was surprised, I didn't think, I didn't expect, I didn't know at first, I didn't understand, I didn't have adequate, I lacked*
10. Expressions of challenge, e.g., *I felt challenged, I was under-prepared, I didn't know how, I wasn't sure, I wasn't comfortable, I felt inadequate, I felt uncertain, I was scared/frightened, I was overwhelmed, it was difficult/hard*
11. Verbs that show pausing, e.g., *I stopped, I paused, this made me stop, I thought, I reflected*
12. Expressions about applying theory to practice, e.g., *I could see how this worked, I learned how to apply, I realized that there are different ways of doing something, what we were taught is not how they do things here*

5.2 Define Formal Rhetorical Patterns

The informal rubrics provide practical linguistic guidance to reflective writing by listing expressions "ready to use," and by associating them with higher-level reflective moves that constitute the rhetorical elements of deep reflective writing: *the description of the relevant context* (Table 1; rubrics 1, 7, and 8), *being specific* (rubrics 2 and 3), *the description of capabilities* (4 and 5), *insights concerning change or*

(2017). Towards reflective writing analytics: Rationale, methodology, and preliminary results. *Journal of Learning Analytics*, *4*(1), 58–84. http://dx.doi.org/10.18608/jla.2017.41.5

shift (4, 6, and 8) and *analysis* (9 and 10). We have set the goal of highlighting the sentences conveying the reflective moves and labelling them according to the rubrics. We consider restricting the analysis to sentences as a first step to reflective writing analytics. Obviously, reflective moves may be accomplished over multiple sentences, but since reflective texts typically include salient sentences that accomplish an entire move, especially shorter reflective texts like the GU corpus, sentences provide a natural testbed for modelling reflective moves. Future work will generalize the method for detecting reflective moves in chains of sentences.

While this list of expressions is an effective support for students (i.e., human language processors) for producing or recognizing reflective moves, it does not provide a sufficient basis for their automated detection for two reasons. First, these expressions only indicate reflection when they are used in appropriate contexts. Consider the following two sentences both of which contain *when,* an indicator for describing context (rubric 1):

> What a mammoth task that turned out to be, partly due to the fact that the motor orientation changes **when** the top of the auger is re-adjusted higher or lower.
>
> I especially love learning about mother–baby interactions, so I was very interested **when** we learned about infant development.

In the first sentence, *when* does not introduce a relevant context for reflection, but in the second sentence it does. This is obvious for a human reader but needs to be modelled computationally.

Second, the list of expressions in the rubrics cannot cover the richness of linguistic expressions that may be used for conveying the reflective moves. Consider the following sentences:

> I am grateful for the practical component that this internship has offered as I feel **I have achieved** a greater **knowledge** of just simply "how things work."
>
> Understanding how the entire company works grants a holistic overview of business operations and often **allows me to understand** the office procedures and processers.
>
> I can already see that **my attitude** towards University **has changed**.

All these sentences describe a *shift in perception* (rubric 6). This move is conveyed by the bold expressions "I have achieved … knowledge," "allows me to understand," and "my attitude … has changed," none of which is a single verb like the examples in the list, and it would be difficult to include them in a list of stereotypical expressions due to their compositional nature. Thus, using merely the expressions in the rubrics for detecting the reflective moves would lead to noise and limited coverage. Our approach for modelling the reflective moves as rhetorical patterns seeks to address both issues: it allows consideration of indicator words and expressions only in those cases where they match the

(2017). Towards reflective writing analytics: Rationale, methodology, and preliminary results. *Journal of Learning Analytics*, *4*(1), 58–84. http://dx.doi.org/10.18608/jla.2017.41.5

patterns, and so only the relevant uses are highlighted; and it makes possible the detection of a great variety of expressions that match the patterns, as described next.

It is based on a deeper conceptual analysis framework called concept-matching, which is applied in the "salient sentences" module of XIP mentioned in Section 4.1. In order to account for the myriad forms of expression conveying rhetorical moves, this analysis framework recognizes that all of the sentences conveying the same rhetorical move share a common underlying conceptual pattern. This pattern is represented as a meta-expression constituted by abstract concepts.

The most basic pattern in reflective moves is AUTHOR's REFLECTION. It is instantiated in sentences by any syntactically related pair of words that refer to the concept of AUTHOR and to the concept of REFLECTION — e.g., "I think," "my idea" — or in more complex ways like "the suggestion that I put forward."

In the case of our example, the pattern underlying "shift in perception" contains an AUTHOR element (e.g., I, me, my), an element pertaining to what we will term MENTAL operations or constructs (e.g., knowledge, understand, attitude), and an element pertaining to CHANGE (e.g., achieved, allows me, changed), which together compose the meta-expression AUTHOR's MENTAL CHANGE.

To implement reflective patterns in XIP, we have added lexicons to the parser, which are lists of words and expressions that can instantiate the various concepts that constitute the meta-expressions modelling reflective moves. These lexicons are taken partly from the analytical writing rhetorical parser previously developed, partly from the rubrics, and partly from the corpora and various synonym lists. The lexicons are evolving using AWA: as new words come up, they can be added to enlarge the coverage of the analysis. Since the parser performs deep dependency analysis, we could develop rules that identify the sentences where the instantiations of the meta-expressions are syntactically related. Figure 2 illustrates the instantiations of the meta-expression as syntactically related words in the three example sentences.

As can be seen, the XIP reflective move categories use the examples in the rubrics as a basis for developing the meta-expressions. Altogether, we implemented the following categories based on the rubrics: *Setting Context*, *Specific Reflection*, *Capability*, *Shift in Perception*, and *Analytical*. The patterns for the last class have been imported from the analytical writing parser (Knight et al., in press).

Our estimation is that it took the XIP analyst five person days' effort to define and conduct preliminary testing of these new sentence types, with a day then needed to update AWA to handle the new XIP output markup, and render them in the user interface.

(2017). Towards reflective writing analytics: Rationale, methodology, and preliminary results. *Journal of Learning Analytics*, *4*(1), 58–84. http://dx.doi.org/10.18608/jla.2017.41.5

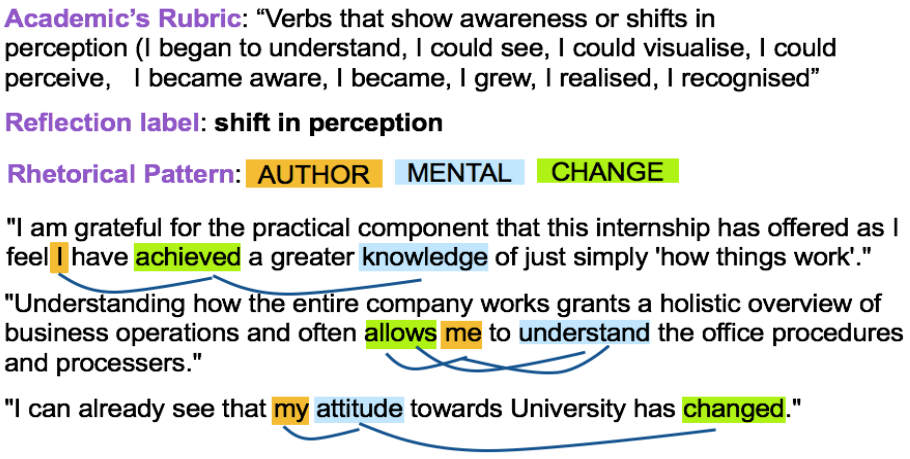

Figure 2: Modelling the elements in a rhetorical pattern associated with reflective writing.

5.3 Independent Reflective Writing Corpus

A corpus of 30 pieces of student reflective writing (containing 382 sentences) was collected and anonymized, selected from university courses that were part of the well-being project at GU described above. The writing was done at the end of the semester and prompted students to reflect on whether the well-being experiences of the course had affected them as a person. The intent was not only to promote reflection on experience — in particular, on the well-being addition to the course — but to also promote self-awareness, interpretation/analysis, and to make connections between personal experience and the theoretical content of the course.

Academic staff and linguistics graduate students coded each writing submission as shallow (surface-level) or deep reflection, as well as whether the reflection extended beyond the personal self to the realm of domain or world (typically expressed as the academic discipline and the student's future role as a professional). Inspired by Carol Rodgers' (2002b) reflective cycle, a coding scheme was developed that distinguished deep from surface-level reflections as those that go beyond description of experience to provide details about how the experience affected or changed a student's way of thinking. The deepest reflections on this spectrum achieve a level of analysis that goes beyond the experience of the self into the realm of the theoretical related to the content of the class (the domain).[5] Sentence-level highlighting was used to identify evidence in support of the code assigned to the reflective piece overall. Coding consisted of trial and revision of rating rubrics, independent coding, subsequent discussion, and finally shared agreement upon coding.

[5] See also Moon (2004) for an extended exploration of levels of reflection.

(2017). Towards reflective writing analytics: Rationale, methodology, and preliminary results. *Journal of Learning Analytics*, *4*(1), 58–84. http://dx.doi.org/10.18608/jla.2017.41.5

An example of a shallow reflection sentence is this: *"I learned so much in this class that I will apply in my life."* Even though this student implies a lot of learning occurred, s/he does not go into detail and describe the learning or the application to life. In contrast, a student who goes into more depth writes, *"So, this course really opened my eyes to some new issues that I had not been aware of before and even to some of the problematic ways I have been taught about my own identity."*

The GU team coded the corpus holistically at the student writing product level, independent of any knowledge of the underlying formal rhetorical patterns modelled in the parser. In this sense, they were coding "freely" as educators, and not with the aim of testing the parser.

6 COMPARISON OF ANNOTATED GU CORPUS WITH PARSER OUTPUT

We now describe the methodology by which we compared the parser output with the GU corpus. In a typical classifier evaluation study, the development corpus (used to design the classifier) and the test corpus are annotated using the same method, which serves as the basis for quality metrics. However, since we had no UTS (or any other) corpus at our disposal that was annotated according to the development rubrics, we used the GU corpus as a proxy and first testbed. The results should therefore be appraised in this context — we are in fact evaluating the degree to which the features of reflective writing identified by the UTS ALL expert, and implemented in XIP, share features in common with those judged to be important, completely independently, at GU. This approach, driven by pragmatic development factors, clearly introduces more complexity and scope for failure. We now describe two iterations.

6.1 Results (First Iteration)

The results assess the degree of overlap between human and machine classification, which provide a metric to quantify progress, and test for improvements/degradations as we iterated. However, we emphasize that these should not be considered a complete evaluation of the system, since classification metrics indicate a system's performance for systems that are clearly comparable with some generally accepted gold standard. Such a gold standard is not at our disposal for our reflective writing rubrics at this point. Moreover, we are not conducting information retrieval or NLP research as computer scientists seeking to improve an algorithm as an end in itself. This is a learning analytics application with intensely pragmatic criteria for success: does automatic feedback improve the current situation? Thus, a more comprehensive evaluation should draw also on evidence from systematic user studies and authentic deployments to assess 1) the effectiveness of the tool in giving actionable feedback, and 2) academic and student reactions to automated feedback that goes well beyond the normal spelling, grammar, and plagiarism checking tools in current usage. Such work is in progress, and will be reported in future publications (Knight et al., in press).

(2017). Towards reflective writing analytics: Rationale, methodology, and preliminary results. *Journal of Learning Analytics*, *4*(1), 58–84. http://dx.doi.org/10.18608/jla.2017.41.5

The output of the automated classification performed by the first version of the parser was compared with the independently annotated GU corpus. For facilitating the comparison, we entered the result of the comparison into a confusion matrix:

TP (true positive) = a sentence labelled as reflective by both the parser and the human analyst

TN (true negative) = a sentence not labelled as reflective by either the parser or the human analyst

FP (false positive) = a sentence labelled as reflective by the parser, but not by the human analyst

FN (false negative) = a sentence labelled as reflective by the human analyst, but not by the parser

The confusion matrix from this evaluation is shown in Table 2, together with the well-established metrics in classification methodology for *Precision, Recall, Accuracy,* and the commonly used aggregate indicator *F1*.

Table 2: Results of the First Evaluation

		ANALYSTS	
		Reflective	Unreflective
XIP	Reflective	**TP:** 35	**FP:** 45
	Unreflective	**FN:** 55	**TN:** 247
Precision	0.438	*P=TP/TP+FP*	
Recall	0.389	*R=TP/TP+FN*	
Accuracy	0.738	*A=(TP+TN)/(TP+FP+FN+TN)*	
F1	**0.412**	*F=2PR/(P+R)*	

Considering the fact that XIP's development and the GU evaluation were entirely independent, these results were promising. Let us take a closer look at the false negatives and the false positives. Regarding false negatives, we identified three types of sentences. The first type contained elements that corresponded to the established patterns, but the words were missing from the reflective lexicon that the parser was using. In this case, adding the words to the lexicon solved the problem. For example, the following sentence was not recognized as conveying a SHIFT due to the lack of the word "realize" in the lexicon:

> Over the past year **I** have come to **realize** that many of my close friends seek support and counseling through campus support and outside healthcare providers.

Once the word is added, the pattern AUTHOR SHIFT is recognized in the XIP dependency SUBJ-N(realize,I), meaning that "I" is the normalized subject of "realize."

(2017). Towards reflective writing analytics: Rationale, methodology, and preliminary results. *Journal of Learning Analytics*, *4*(1), 58–84. http://dx.doi.org/10.18608/jla.2017.41.5

The second type of false negative contained sentences where no reflective pattern was found. This is the case in the following sentence highlighted by the human annotator:

> When I walk into a lecture hall, I look for a familiar face, perhaps one that I met during [course name].

The human analyst identified this sentence as the last of four sentences that together were representing a reflection on the student's experience in the course, which had resulted in a change that s/he carried into other settings:

> The environment was welcoming and comfortable, so it was much easier to discuss matters such as those in a classroom with other students and a professor when normally conversations of that nature would take place among friends. [Course name] cultivated an environment where we were able to learn from each other and build off of other ideas. Looking back on the semester, I don't think I could have felt as comfortable and at ease as I do now without this class. When I walk into a lecture hall, I look for a familiar face, perhaps one that I met during [course name].

The semantics of "shift" in the parser, however, includes a shift in learning or reflection, which is not the case in the last sentence. This is why it is not selected. In this case, the XIP category did not cover the analyst's category, which also included a shift in behaviour. This may also be a case where the human annotator was coding the meaningful details that followed the reflective set-up identified by the parser (see "false positives," below). The parser had selected as reflective the third sentence in this example, whereas the human focused on the result of the reflection, which in this case appears in a new sentence. If these sentences had been connected by a semi-colon or woven together, the whole sentence would have been chosen by the parser to include this content. This is a limitation of the sentence-level analysis.

The third type of false negative led us to add two new patterns that were not conveyed by the reflective rubrics: sentences that describe other people's point of view and reflections about the class, as in the following sentence:

> **For some**, it was described as less pressuring and time constrained than high school, while **others felt like** college made them give up some free time they may have had in the past.

Concerning the false positives, the annotators considered that several of them could indeed be annotated as deeper reflections, but they were not highlighted because the same idea had been expressed earlier in the essay (see description of annotators' feedback below). Some other false positives were the result of too loose an implementation of the patterns. For example, the *Capability* pattern whose rubric is "Expressions of increased confidence or ability (*am more confident, am now able, feel/am comfortable, can plan, can utilize, can develop a strategy*)" erroneously classified the following sentence:

(2017). Towards reflective writing analytics: Rationale, methodology, and preliminary results. *Journal of Learning Analytics*, *4*(1), 58–84. http://dx.doi.org/10.18608/jla.2017.41.5

> Through different people's reactions to this situation **I was able to learn** about the different ways people would solve her situation and whether or not they really felt all that bad for her deviance.

The solution to this kind of false positive is adding restrictive rules that exclude them, even though there was agreement that this is an important category. In this case, we decided to exclude the *Capability* type temporarily, because of the time constraints for new rule development.

A major result of this first iteration was that it gave rise to more subtle rules for filtering out shallow reflections from deeper ones. Since the human annotation focused on high quality reflections, some of the false positives revealed cases when the sentence did contain a reflective pattern, but the reflection itself had a shallow content. The following sentence is an example:

> **I found** that the way to present your attributes and skills **is** essential.

This sentence does describe the author's attitude, but fails to go beyond the self and explore reasons for the attitude or any links with some deeper, theoretical consideration expected in deep reflection (cf. Section 5.3). As an experiment, we established some basic rules that might reflect some linguistic indications of shallow reflection. These rules match sentences without some complexity of linguistic structure like subordination or only satisfy a pattern of "specific" reflection without any other reflective pattern, like the sentence above.

Taken together, the first iteration allowed us to make significant improvements in the system, as evidenced in the second iteration. New XIP sentence categories for *Superficial* (shallow) reflections were added. Not discussed in this paper were additional categories where the students reflect on how their experiences relate to what is being learned in formal *Class,* and deeper reflections that go beyond expressing personal views about a context and take into account the *Viewpoints* of other stakeholders (see Figure 1: user interface).

6.2 Results (Second Iteration)

In developing the second version, we took into account the errors and missed sentences in the first iteration: we expanded the lexicon, disambiguated some words, and introduced new sentence labels. Table 3 shows some improvement of the results on the corpus of 30 annotated texts. As the table shows, adding new words and filtering out surface reflection, as expected, significantly improves recall, and somewhat improves precision.

After this preliminary testing, we obtained an expanded corpus of annotated extracts from the Georgetown University team containing 312 extracts and 2366 sentences. Table 4 shows the results of this evaluation compared to Table 3. Clearly, there is much scope to improve performance in terms of these metrics (see Discussion on future sources of other evaluation data). However, accuracy did not decrease significantly, which is promising since the new evaluation corpus had almost ten times as many

(2017). Towards reflective writing analytics: Rationale, methodology, and preliminary results. *Journal of Learning Analytics*, *4*(1), 58–84. http://dx.doi.org/10.18608/jla.2017.41.5

sentences as the first, which increases the number of potential new words that might not have been recognized by XIP. As for the degradations in other indices, we discuss this in Section 7.2.

Table 3: Results of Second Test*

		ANALYSTS	
		Reflective	Unreflective
XIP	Selected	**TP:** 53	**FP:** 51
	Unselected	**FN:** 32	**TN:** 278
Precision	0.509 **(+0.071)**	*P=TP/TP+FP*	
Recall	0.623 **(+0.234)**	*R=TP/TP+FN*	
Accuracy	0.799 **(+0.061)**	*A=(TP+TN)/(TP+FP+FN+TN)*	
F1	**0.560 (+0.148)**	*F=2PR/(P+R)*	

* Brackets show improvement on the first iteration in Table 2.

Table 4: Test Results on a Larger Corpus**

		ANALYSTS	
		Reflective	Unreflective
XIP	Selected	**TP:** 129	**FP:** 494
	Unselected	**FN:** 219	**TN:** 1524
Precision	0.207 **(–0.302)**	*P=TP/TP+FP*	
Recall	0.37 **(–0.253)**	*R=TP/TP+FN*	
Accuracy	0.698 **(–0.101)**	*A=(TP+TN)/(TP+FP+FN+TN)*	
F1	**0.266 (–0.294)**	*F=2PR/(P+R)*	

** Brackets show the degradation with respect to Table 3.

6.2.1 Classifying shallow reflections

We have two preliminary, indirect indications that the parser could detect shallow reflections. Firstly, we compared the ratio of sentences labelled as shallow reflections in a good and a poor UTS engineering report (recall these are sizeable, 40–50 pages), and found that in the good report 26% of the reflective sentences were annotated as shallow reflection against 48% in the poor report, almost twice as many. Although just one case, this corresponds to the direction one would hope for.

Secondly and more robustly, we compared sentences labelled as shallow reflections by the parser with the human annotations in the entire annotated GU corpus of reflections. Of the 209 reflections marked as shallow by the parser, 49 were annotated by the human annotators as deep reflections, with the remaining 160 uncoded (i.e., by implication, shallow). Although more rigorous evaluation is necessary, these two tests indicate that the shallow reflection classifier may add value to the analysis.

(2017). Towards reflective writing analytics: Rationale, methodology, and preliminary results. *Journal of Learning Analytics*, *4*(1), 58–84. http://dx.doi.org/10.18608/jla.2017.41.5

7 DISCUSSION AND FUTURE WORK

7.1 A Closer Look at False Positives

We have shown that from the first to second design iterations, we were able to demonstrate immediate, albeit minor, improvements by making small changes of several different sorts to XIP in light of feedback from the Georgetown University academics who performed the hand coding. As the academics explained when they commented on AWA's output, they approached the coding of the writing (which was almost all reflective to some degree) in a more holistic manner than the exhaustive sentence-by-sentence procedure used by XIP (emphasis added). Moreover, they set the bar high in their criteria:

> First, although we coded sentences, we were fairly focused on assigning a code to the overall essay, so *we were focused more at the student level rather than the sentence level*. Our approach meant that in practice we were highlighting sentences that were the *most* reflective, or had the *most evidence*. We did not always comprehensively highlight. Many of the sentences [XIP] found are contained in essays we had coded as reflective overall, but we had left out that particular sentence.
>
> Second, because our initial coding and the nature of the assignment indicated that we had a corpus that was largely reflective (and we have evidence that 99% of the "cases" of these student essays had self-reflection) we left uncoded reflection that was "merely" what we would have called surface-level self-reflection. We *only coded sentences that either pushed the envelope on the depth scale or pushed from the self to be reflecting on domain or world*. In essence, we agree that there are many surface-level self-reflective sentences in here that we didn't code. But your parser found a lot of those!

In looking at the false positives, the human annotators had these additional observations about the types of patterns that seemed to generate false positives. One FP pattern seemed to be where the parser was correctly recognizing sentence-level reflection, but the annotator had disregarded that sentence as deeply reflective because it was set in the context of an extremely short piece of writing, typically containing only two sentences. If an instructor is looking for meaningful student reflection, it typically does not occur with the amount of desired detail in two sentences. For example, the parser identified the second sentence of this two-sentence essay as reflective, but the human coder had ignored it because of its lack of detail or explication.

> Before I came to this class I had never really thought much about gender and what it means or that it is something that is fluid. Taking this course was completely eye opening and really made me think about things I have never had the chance to think about.

Along similar lines, the human annotators had originally approached their coding in taking a whole-essay or whole-text approach. In this approach, essay entries such as the one above would not have qualified

(2017). Towards reflective writing analytics: Rationale, methodology, and preliminary results. *Journal of Learning Analytics*, *4*(1), 58–84. http://dx.doi.org/10.18608/jla.2017.41.5

for deep reflection because of lack of detail. For the purposes of comparing with the sentences highlighted by the parser, the annotators highlighted sentence-level evidence for the more holistic approach, and probably did so less systematically than the parser.

In another FP pattern, the annotator interpreted a sentence as descriptive whereas the parser highlighted it as reflective. This may have been because the annotators were looking specifically for a *personal self-reflection* where the student was integrating content with their own personal experience and thoughts. The parser, on the other hand, selected sentences where the student was reflecting generally on the course environment for everyone or the mode of teaching as being effective.

In a third pattern, the human annotator highlighted a sentence following one that the parser selected. During analysis, it became clear that the parser was identifying the reflective "set-up," and the human was focusing on the meaningful content that then followed. The annotators were not trained in recognizing particular reflective moves, nor were they coding for these moves. When reviewing AWA's output, it was clear to the GU analysts that they were often noticing the meaningful description, which came after the linguistic reflective construction. When these were separated into two different sentences, the coding between human and parser did not overlap, even though, taken as a whole, they were both finding the same passage.

All efforts to develop and validate writing analytics must navigate this kind of difference in the way that people and machines make sense of a text. These specific comments were encouraging in the sense that the academics had set a high threshold for their highlighting of deeper reflection. Perhaps in order to be truly useful to instructors for assessing and to students for feedback and improvement, an automated parser would ideally need to incorporate a two-stage process. The first would involve identifying sentence-level reflective moves, and the second would re-evaluate the analysis of the selected sentences within the context of the whole piece, or the moves being made in that piece.

7.2 Future Work

Several considerations have been the focus of subsequent developments to what is reported in this paper. Firstly, the *user interface* has not been the focus of this paper, but we are developing this beyond the example shown. We are considering a range of design options that will go beyond highlighting sentences in order to improve the specificity of the feedback (cf. Hulsmann & van der Floodt, 2015) to help the user identify what strategies they might need to consider in order to address weaknesses. Secondly, XIP currently operates on *single sentences*, which is clearly a limitation, since we do not write or assess reflective pieces as a series of isolated sentences. Thirdly, there is of course much more to the quality of a text than just the reflective moves. We see XIP as just one of a potential suite of text analytics services, which could expand to include other metrics. We now have a *text analytics infrastructure* (TAP) to enable AWA to call on a scalable, distributed analytics infrastructure. Our most recent design iteration reports the development of a more sophisticated framework for reflective

(2017). Towards reflective writing analytics: Rationale, methodology, and preliminary results. *Journal of Learning Analytics*, *4*(1), 58–84. http://dx.doi.org/10.18608/jla.2017.41.5

writing to inform the analytics, corresponding extensions to NLP modelling that use TAP, and user interface feedback that extends beyond the sentence level, accompanied by new student usage data (Gibson et al., 2017).

TAP should permit us to orchestrate different analytics workflows to investigate the relevance of other patterns in a corpus; for example, do assignments with high grades have distinctively different sequences/locations of sentence types? Preliminary steps on this front are reported by Knight, Maldonado-Martinez, Gibson, and Buckingham Shum (2017).

7.3 On the Risks of Gaming the System

A justified concern around machine analysis of writing is that students seek to reverse engineer the features of interest to the parser, and then reproduce them in a meaningless way. However, we do not consider this a realistic danger since AWA is not being used for summative grading purposes, but to provide rapid formative feedback by highlighting and tagging potentially relevant reflective elements. The students are thus only fooling themselves, and in other contexts when we give AWA briefings to students, we emphasize that the machine will make mistakes, and that final grade is a function of more factors than the mere presence of the right rhetorical features. The relevance of the reflection should take into account the entire content of the sentence, and as noted, the meaning at the paragraph or even whole document level, which remains the province of human interpretation. The opening line of the feedback page reminds users: *"AWA does not of course know if it is beautifully crafted nonsense — you must decide that."*

Used formatively, therefore, there should be no "secret" about sharing with students the linguistic features driving AWA — quite the opposite. The *rubrics* that are the foundation of the automated analysis are shared with students, providing the language and exemplars for reflection that are so often missing from their experience. Moreover, AWA's output will use terminology consistent with the rubrics. According to this approach, students should be encouraged to argue with the machine when they disagree with the feedback. Assuming there is an acceptable signal-to-noise ratio, this is exactly the higher level of discourse that we want to provoke. Academics have often proposed to us that they could envisage productive collaborative activities in which pairs of students use their AWA reports as a springboard for discussion with each other. Initial student feedback from Knight et al. (in press) suggests that (compared to feedback from a tutor), for some students the machine's dispassionate analysis can make it easier to accept and reflect on poor writing — but equally, students would be encouraged to "push back," and thus develop the kind of critical mind about machine intelligence that is now required as a lifelong learning capacity (cf. Buckingham Shum et al., 2016). The testing of learning/instructional design patterns, in which writing analytics are thoroughly integrated and therefore meaningful to students and staff, is now a critical area for development.

(2017). Towards reflective writing analytics: Rationale, methodology, and preliminary results. *Journal of Learning Analytics*, *4*(1), 58–84. http://dx.doi.org/10.18608/jla.2017.41.5

7.4 Participatory Design to Build Trust

We have described a methodology for rapid prototyping as a way to build trust among key stakeholders. While newcomers to writing analytics can understand in principle what the potential of NLP is, it is only when the ALL expert and the academics at UTS and GU could see for themselves how AWA behaved that this potential became tangible. Central to this dynamic is good communication, mediated by the learning analytics R&D team as brokers and designers. Central to the mainstreaming of writing analytics tools is trust among the key stakeholders.

This dialogue was conducted through a mix of synchronous and asynchronous exchanges across three countries. The important quality is reciprocity, such that all parties are learning from each other. A key relationship in question is whether the ALL expert trusts that her work is being translated with *transparency* (she understands the process) and *integrity* into the XIP rhetorical patterns (the results match her judgements). The GU academics were not involved in the design of the initial patterns, but gave feedback on *integrity*, which led to conversations about how XIP worked, and to changes being made.

Trust is built through reciprocity, which in learning analytics design means ultimately that *you feel you can influence the code*. While the core team can of course directly change AWA, we envisage offering ways for users 1) to give direct feedback to AWA on the usefulness of the sentences it is highlighting, and 2) to edit the lexicon so that generic and discipline-specific terminology causing false positives and negatives can be reduced. We can expand the circle of users able to exert control over their tools by learning from the "end-user development" community who have studied the ecosystems that evolve around software tools that permit different kinds of end-users to modify the application's behaviour to differing degrees, and the different user interfaces and exchange mechanisms that enable this (Burnett & Scaffidi, 2011; MacLean, Carter, Lövstrand, & Moran, 1990).

In principle, this co-design approach should generalize to other contexts, and to other kinds of analytics, depending on the quality of the common ground and reciprocity that can be established.

7.5 Piloting with Students

This paper has documented the preliminary steps to validating a writing analytics application. Firstly, we needed to build the confidence of reflective writing experts that the XIP parser has a classification scheme based in sound pedagogy and scholarship. Secondly, we wanted to quantify performance quality, and although there is much room for improvement in this extremely challenging domain, we are encouraged that the parser was able to produce promising results on an unseen corpus, sourced and annotated independently from the AWA team.

The user interface went through rapid prototyping with the ALL expert (and many other UTS academics testing it for their texts) using think-aloud walkthroughs. The resulting design served as a sufficiently

(2017). Towards reflective writing analytics: Rationale, methodology, and preliminary results. *Journal of Learning Analytics*, *4*(1), 58–84. http://dx.doi.org/10.18608/jla.2017.41.5

intuitive rendering that the GU team had no difficulty in understanding how to make sense of it when reviewing and critiquing output. More detailed usability evaluations will help refine the design as we move to student trials.

Once the academics are satisfied that AWA adds more value than distraction, and that the user experience is good enough, the next step is to introduce students to it. Following the approach taken in piloting AWA for analytical (as opposed to reflective) academic legal writing (Knight et al., in press), our approach is to work closely with academics to ensure that AWA is integrated with the curriculum's learning design and assessment regime, to maximize the meaningfulness of testing AWA. This requires that the language used in AWA is aligned with the way that reflection and reflective writing is taught (which has inevitable variations depending on discipline and level of student). Ideally, we will be able to devise a modelling approach that is comprehensible and acceptable to different academics. The data to inform AWA's evaluation will span system logs to reveal usage patterns, student surveys for user feedback, and the quality of reflective writing. Gibson et al. (2017) describe how this process was followed in the context of reflective writing among pharmacy students, building on the work in this paper.

8 CONCLUSION

We have introduced the distinctive purposes of reflective writing as practised in educational contexts for decades, as well as the complexities this creates for teaching, learning, and assessment. This has been the subject of active research independent of, and preceding, the emergence of learning analytics. Recognizing and understanding this evidence base sets the context for any learning analytics design effort.

We ask our students to make their thinking visible in their writing, but to do so they must understand what this looks like. Beyond exposing them to principles and examples, the real learning occurs when they receive coaching on their own writing, but this is time-consuming to provide, and is in fact a task that many academics find challenging. Given the challenges of teaching, learning, and assessing academic reflective writing, we have identified the potential of providing instant formative feedback on draft writing — student work that would otherwise receive no feedback due to the limited availability of educator time. A writing analytics tool such as AWA goes beyond rubrics that make explicit the important features of this genre of writing *in general* by highlighting the linguistic forms it finds *in the student's own text instantaneously*.

The academics engaged with the AWA team do not feel threatened by this kind of machine intelligence, appreciating its potential to address their limited resources. AWA shows potential as a vehicle for codifying informal rubrics for academic reflective writing in a form that is accessible to academics, tutors, and students. If AWA fulfills its promise, we are moving to a scenario of being able to offer 24/7 formative feedback to learners, on their own drafts or any other text they choose to reflect on.

(2017). Towards reflective writing analytics: Rationale, methodology, and preliminary results. *Journal of Learning Analytics*, *4*(1), 58–84. http://dx.doi.org/10.18608/jla.2017.41.5

With the growing availability of writing analytics products and prototypes, new opportunities present themselves to validate effective learning designs that incorporate automated feedback into student learning. This feedback could be socialized for instance, serving as a provocation for discussion among peers and/or tutors on the quality of good reflective writing.

We go further to propose that the possibility of formally modelling the principles of reflective writing in a parser opens interesting avenues for the writing research community. Tools such as AWA provide an empirical tool for investigating to what extent different genres of reflection share linguistic features in common, and in what ways they differ. The ease or difficulty with which one can formalize a particular theory or model of writing is itself an instructive process to undertake; the dilemmas one faces, and the decisions one makes, can be enlightening, perhaps revealing inconsistencies or gaps in the theory.

We close by returning to the drivers that make learners' reflections a topic of such interest. There is a growing movement of educators calling for assessment regimes (and by extension formal, validated metrics) that value a holistic conception of learning, preparing the learner for the complexities of the workplace and society more broadly. See the recent analysis of the challenges facing liberal arts in the digital era by Bass and Eynon (2016), and the recent volume dedicated to analytics for building 21st century competencies (Buckingham Shum & Deakin Crick, 2016). Reflective writing analytics is a form of learning analytic for advancing this agenda, with its emphasis on integrating theory with practice, honouring both the mind and the emotions, and valuing not only what the learner can display in their mastery of the material, but what they can articulate as challenging and uncertain.

ACKNOWLEDGEMENTS

We gratefully acknowledge the University of Technology Sydney for the Vice Chancellor's strategic grant that funded the *Authentic Assessment Analytics for Reflection (A3R)* project.

REFERENCES

Aït-Mokhtar, S., Chanod, J-P., & Roux, C. (2002). Robustness beyond shallowness: Incremental deep parsing. *Natural Language Engineering*, *8*(2/3), 121–144. http://dx.doi.org/10.1017/S1351324902002887

Bass, R., & Eynon, B. (2016). *Open and integrative: Designing liberal education for the new digital ecosystem*. Association of American Colleges and Universities. https://www.aacu.org/publications-research/publications/open-and-integrative-designing-liberal-education-new-digital

Boud, D., Keogh, R., & Walker, D. (1985). *Reflection: Turning experience into learning.* London: Routledge, Abingdon, Oxon.

Boud, D., & Walker, D. (1998). Promoting reflection in professional courses: The challenge of context. *Studies in Higher Education*, *23*(2), 191–206. http://dx.doi.org/10.1080/03075079812331380384

(2017). Towards reflective writing analytics: Rationale, methodology, and preliminary results. *Journal of Learning Analytics*, *4*(1), 58–84. http://dx.doi.org/10.18608/jla.2017.41.5

Brun, C., & Hagège, C. (2003). Normalization and paraphrasing using symbolic methods. *Proceedings of the 2nd International Workshop on Paraphrasing* (PARAPHRASE '03), Vol. 16, 11 July 2003, Sapporo, Japan (pp. 41–48). Stroudsburg, PA: Association for Computational Linguistics. http://dx.doi.org/10.3115/1118984.1118990

Brun, C., & Hagège, C. (2004). Intertwining deep syntactic processing and named entity detection. *Advances in natural language processing* (pp. 195–206). Berlin: Springer. http://dx.doi.org/10.1007/978-3-540-30228-5_18

Brun, C., Popa, D. N., & Roux, C. (2014). XRCE: Hybrid classification for aspect-based sentiment analysis. *Proceedings of the 8th International Workshop on Semantic Evaluation* (SemEval 2014), 23–24 August 2014, Dublin, Ireland (pp. 838–842). Association for Computational Linguistics. http://www.aclweb.org/anthology/S14-2149

Buckingham Shum, S., & Deakin Crick, R. (2016). Learning analytics for 21st century competencies. *Journal of Learning Analytics, 3*(2), 6–21. http://dx.doi.org/10.18608/jla.2016.32.2

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. *Proceedings of the 6th International Conference on Learning Analytics and Knowledge* (LAK '16), 25–29 April 2016, Edinburgh, UK (pp. 213–222). New York: ACM. http://dx.doi.org/10.1145/2883851.2883955

Burnett, M. M., & Scaffidi, C. (2011). End-user development. In M. Soegaard & R. F. Dam (Eds.), *The encyclopedia of human–computer interaction,* 2nd ed. Interaction Design Foundation. https://www.interaction-design.org/literature/book/the-encyclopedia-of-human-computer-interaction-2nd-ed

De Liddo, A., Sándor. Á., & Buckingham Shum, S. (2012). Contested collective intelligence: Rationale, technologies, and a human–machine annotation study. *Computer Supported Cooperative Work, 21*(4–5), 417–448. http://dx.doi.org/10.1007/s10606-011-9155-x

Gibson, A., & Kitto, K. (2015). Analysing reflective text for learning analytics: An approach using anomaly recontextualisation. *Proceedings of the 5th International Conference on Learning Analytics and Knowledge* (LAK '15), 16–20 March 2015, Poughkeepsie, NY, USA (pp. 275–279). New York: ACM. http://dx.doi.org/10.1145/2723576.2723635

Gibson, A., Aitken, A., Sándor, Á., Buckingham Shum, S., Tsingos-Lucas, C., & Knight, S. (2017). Reflective writing analytics for actionable feedback. *Proceedings of the 7th International Conference on Learning Analytics and Knowledge* (LAK '17), 13–17 March 2017, Vancouver, BC, Canada (pp. 153–162). New York: ACM. http://dx.doi.org/10.1145/3027385.3027436

Gibson, A., Kitto, K., & Bruza, P. (2016). Towards the discovery of learner metacognition from reflective writing. *Journal of Learning Analytics, 3*(2), 22–36. http://dx.doi.org/10.18608/jla.2016.32.3

Hatton N., & Smith, D. (1995, January). Reflection in teacher education: Towards definition and implementation. *Teaching & Teacher Education, 11*(1), 33–49. http://dx.doi.org/10.1016/0742-051X(94)00012-U

(2017). Towards reflective writing analytics: Rationale, methodology, and preliminary results. *Journal of Learning Analytics*, *4*(1), 58–84. http://dx.doi.org/10.18608/jla.2017.41.5

Huang, Z., ten Teije, A., van Harmelen, F., & Aït-Mokhtar, S. (2014). Semantic representation of evidence-based clinical guidelines. *Proceedings of the 6th International Workshop on Knowledge Representation for Health Care* (KR4HC 2014), 21 July 2014, Vienna, Austria (pp. 78–94). http://dx.doi.org/10.1007/978-3-319-13281-5_6

Hulsman, R. L., & van der Vloodt, J. (2015). Self-evaluation and peer-feedback of medical students' communication skills using a web-based video annotation system. Exploring content and specificity. *Patient Education and Counseling*, *98*(3), 356–363. http://dx.doi.org/10.1016/j.pec.2014.11.007

Knight, S., Buckingham Shum, S., Ryan, P., Sándor, Á., & Wang, X. (in press). Designing academic writing analytics for civil law student self-assessment. *International Journal of Artificial Intelligence in Education* (Special Issue on Multidisciplinary Approaches to Reading and Writing Integrated with Disciplinary Education, Eds. D. McNamara, S. Muresan, R. Passonneau & D. Perin). http://dx.doi.org/10.1007/s40593-016-0121-0

Knight, S., Martinez-Maldonado, R., Gibson, A., & Buckingham-Shum, S. (2017). Towards mining sequences and dispersion of rhetorical moves in student written texts. *Proceedings of the 7th International Conference on Learning Analytics and Knowledge* (LAK '17), 13–17 March 2017, Vancouver, BC, Canada (pp. 228–232). New York: ACM. http://dx.doi.org/10.1145/3027385.3027433

Lisacek, F., Chichester, C., Kaplan, A., & Sandor, Á. (2005). Discovering paradigm shift patterns in biomedical abstracts: Application to neurodegenerative diseases. *Proceedings of the 1st International Symposium on Semantic Mining in Biomedicine* (SMBM), 11–13 April 2005, Cambridge, United Kingdom (pp. 41–50). http://www.xrce.xerox.com/Research-Development/Publications/2005-006

MacLean, M., Carter, K., Lövstrand, L., & Moran, T. (1990). User-tailorable systems: Pressing the issues with buttons. *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (CHI '90), 1–5 April 1990, Seattle, WA, USA (pp. 175–182). New York: ACM. http://dx.doi.org/10.1145/97243.97271

Moon, J. (2004). *A handbook of reflective and experiential learning: Theory and practice.* London: RoutledgeFalmer.

Moon, J. (2010). Reflective learning workshop (Handout 10/07), University of Worcester, UK. http://worc.ac.uk/edu/documents/Jenny_Moon_RefLearnlong07.doc

Reidsema, C., Goldsmith, R., & Mort, P. (2010). Writing to learn: Reflective practice in engineering design. *Proceedings of the 9th Annual ASEE Global Colloquium* (ASEE 2010), 18–21 October 2010, Singapore. American Society for Engineering Education. https://www.academia.edu/501660/Writing_to_Learn_Reflective_Practice_in_Engineering_Design

Rodgers, C. (2002a). Defining reflection: Another look at John Dewey and reflective thinking. *Teachers College Record, 104*(4), 842–866. http://www.tcrecord.org/content.asp?contentid=10890

Rodgers, C. (2002b). Voices inside schools. *Harvard Educational Review*, *72*(2), 230–254.

(2017). Towards reflective writing analytics: Rationale, methodology, and preliminary results. *Journal of Learning Analytics*, *4*(1), 58–84. http://dx.doi.org/10.18608/jla.2017.41.5

Ryan, M. (2010). The 4 R's model of reflective thinking, version 1.5. *Developing Reflective Approaches to Writing (DRAW) Project*, Queensland University of Technology. http://www.citewrite.qut.edu.au/write/4Rs-for-students-page1-v1.5.pdf

Ryan, M. (2011). Improving reflective writing in higher education: A social semiotic perspective, *Teaching in Higher Education*, *16*(1), 99–111. Http://dx.doi.org/10.1080/13562517.2010.507311

Ryan, M., & Ryan, M. (2012). Developing a systematic, cross-faculty approach to teaching and assessing reflection in higher education. Australian Government, Office of Learning and Teaching. http://www.olt.gov.au/system/files/resources/PP9_1327_Ryan_report_2012.pdf

Sándor, Á. (2007). Modeling metadiscourse conveying the author's rhetorical strategy in biomedical research abstracts. *Revue Française de Linguistique Appliquée, 12*(2), 97–108. http://www.xrce.xerox.com/Research-Development/Publications/2007-029

Sándor, Á., & Vorndran, A. (2009). Detecting key sentences for automatic assistance in peer reviewing research articles in educational sciences. *Proceedings of the Workshop on Text and Citation Analysis for Scholarly Digital Libraries*, 47th Annual Meeting of the Association for Computational Linguistics, 2–7 August 2009, Singapore. http://www.xrce.xerox.com/Research-Development/Publications/2009-039

Scouller, K. (1998). The influence of assessment methods on students' learning approaches: Multiple choice question examination versus assignment essay. *Higher Education*, *35*(4), 453–472. http://dx.doi.org/10.1023/A:1003196224280

Simsek, D., Buckingham Shum, S., Sándor, Á., De Liddo, A., & Ferguson, R. (2013). XIP dashboard: Visual analytics from automated rhetorical parsing of scientific metadiscourse. 1st International Workshop on Discourse-Centric Learning Analytics (DCLA13), 8 April 2013, Leuven, Belgium. http://oro.open.ac.uk/37391

Simsek, D., Sándor, Á., Buckingham Shum, S., Ferguson, R., De Liddo, A., & Whitelock, D. (2015). Correlations between automated rhetorical analysis and tutors' grades on student essays. *Proceedings of the 5th International Conference on Learning Analytics and Knowledge* (LAK '15), 16–20 March 2015, Poughkeepsie, NY, USA (pp. 355–359). New York: ACM. http://dx.doi.org/10.1145/2723576.2723603

Sumsion, J., & Fleet, A. (1996). Reflection: Can we assess it? Should we assess it? *Assessment & Evaluation in Higher Education*, *21*(2), 121–130. http://dx.doi.org/10.1080/0260293960210202

Ullmann, T. D., Wild, F., & Scott, P. (2012). Comparing automatically detected reflective texts with human judgements. *Proceedings of the 2nd Workshop on Awareness and Reflection in Technology-Enhanced Learning* (AR-TEL '12), 18 September 2013, Saarbrucken, Germany (pp. 101–116). http://ceur-ws.org/Vol-931/paper8.pdf

Webster-Wright, A. (2013). The eye of the storm: A mindful inquiry into reflective practices in higher education. *Reflective Practice, 14*(4), 556–567. http://dx.doi.org/10.1080/14623943.2013.810618