# Reflecting on Reflective Writing Analytics: Assessment Challenges and Iterative Evaluation of a Prototype Tool

Simon Buckingham Shum[1], Ágnes Sándor[2], Rosalie Goldsmith[3],
Xiaolong Wang[1], Randall Bass, Mindy McWilliams[4]

[1] Connected Intelligence Centre
[3] Inst. for Interactive Media in Learning
University of Technology Sydney
Broadway, Ultimo, NSW 2007, AUS
first.lastname@uts.edu.au

[2] Xerox Research Centre Europe
6 chemin Maupertuis
F-38240 Meylan
FRANCE
agnes.sandor@xrce.xerox.com

[4] Georgetown University
37th and O Streets N.W.
Washington D.C. 20057, USA
bassr@georgetown.edu
mcwillie@georgetown.edu

## 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, reflecting 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. This has been through iterative evaluation on an independently human-annotated corpus, showing improvements from the first to second version. We conclude by discussing the reasons why classifying reflective writing has proven complex, and reflect on the design processes enabling work across disciplinary boundaries to develop the prototype to its current state.

## Categories and Subject Descriptors

K.3.1 [**Computers and Education**]: Computer Uses in Education

## General Terms

Algorithms, Design, Experimentation, Human Factors, Measurement

## Keywords

Learning Analytics, Education, Writing Analytics, Reflection, Natural Language Processing, Metadiscourse, Rhetoric

*LAK '16,* April 25-29, 2016, Edinburgh, United Kingdom © 2016 ACM. ISBN 978-1-4503-4190-5/16/04…$15.00
DOI: http://dx.doi.org/10.1145/2883851.2883955

## 1. ACADEMIC REFLECTIVE WRITING

Reflection has long been regarded as a key element in student learning and professional practice in higher education [2, 10, 16, 18]. It can allow students a window into their developing professional identity [15], deepen understanding of key concepts [24], and provide opportunities for lifelong learning [18]. However, it has been so broadly interpreted and implemented in the university curriculum that the concept of reflection has become attenuated [28]. Because of such broad interpretations, defining what is meant by reflection is no easy task [16]. The definition by Boud, Keogh and Walker [2] provides a useful perspective:

> *Reflection is an important human activity in which people recapture their experience, think about it, mull over & evaluate it. It is this working with experience that is important in learning* ([2], 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 [2, 10]. 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 [2, 26].

However, reflective writing is for many students, and educators, a novel genre to compose in, and to assess. We introduce the complexities next (Sec.2), and describe the particular contexts in which we are using reflective writing (Sec.3). We then introduce the technical platform we are developing (Sec.4), before moving to describe the methodology by which we move from rubrics, to rhetorical patterns (Sec.5). Two iterations of the parser are then detailed (Sec.6), before 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 (Sec.7).

## 2. ASSESSMENT CHALLENGES

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 and 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 [10, 26].

Related to this is the need for a shared language to teach and assess reflective writing, as identified by Ryan [18] in a project specifically intended to develop the teaching of reflective writing across a number of disciplines in an Australian university [20] 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 students' assignments.

Boud and Walker put forward the argument that as reflective writing is very different in nature and purpose from analytical academic writing, it should be assessed using criteria that are sensitive to that particular genre ([3], p.194). In their seminal paper on how and whether to assess reflective writing tasks, Sumsion and Fleet 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 ([26], p.124). This is an 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 understand that expressed 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 straightforward. 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 they 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 personalised 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. 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 the light of the evidence of 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 analysed 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 could in principle 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?

# 3. REFLECTIVE WRITING CONTEXTS

## 3.1 Reflective writing for Engineers (UTS)

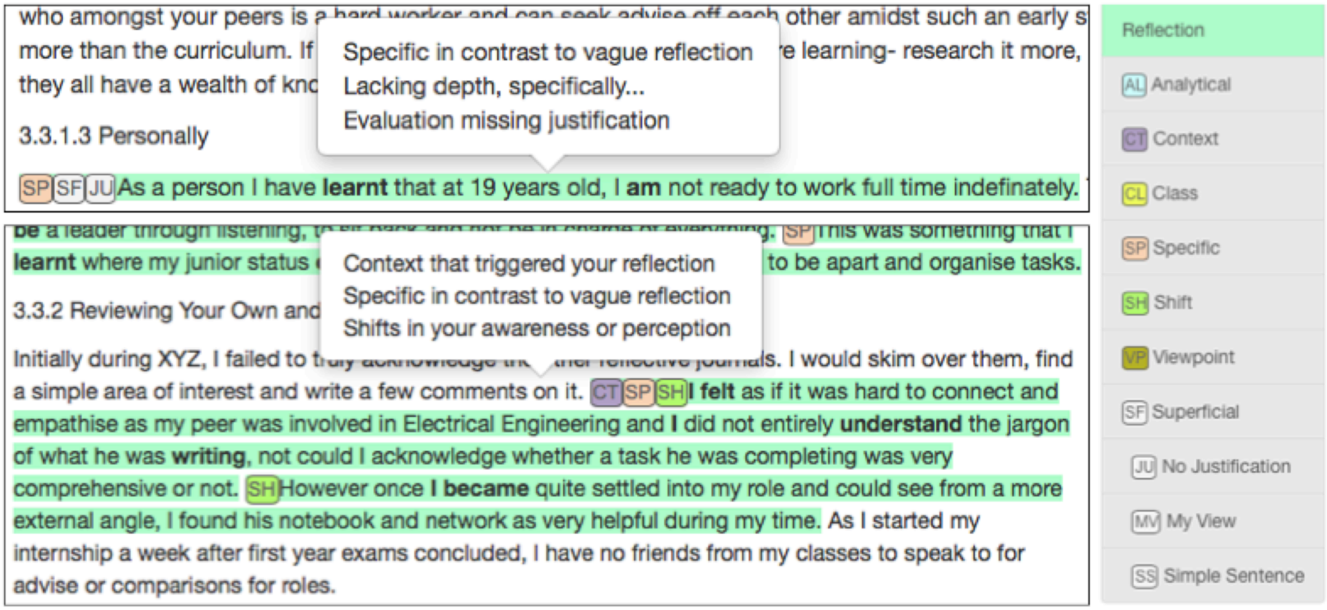
At the University of Technology Sydney, all engineering students in the 4 year degree program undertake two 6-month internships which are part of the practice program. At the completion of each internship students are required to submit a reflective report that details changes in their professional, personal and technical awareness. The cohort size is approximately 200 per semester, with reports expected to be 40-50 pages, and hence 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. 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 this quality? 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

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



who amongst your peers is a hard worker and can seek advise off each other amidst such an early s
more than the curriculum. If
they all have a wealth of knc
e learning- research it more,

Specific in contrast to vague reflection
Lacking depth, specifically...
Evaluation missing justification

3.3.1.3 Personally

SP SF JU As a person I have **learnt** that at 19 years old, I **am** not ready to work full time indefinately.

be a leader through listening, to sit back and not be in charge of everything. SP This was something that I
**learnt** where my junior status
to be apart and organise tasks.

Context that triggered your reflection
Specific in contrast to vague reflection
Shifts in your awareness or perception

3.3.2 Reviewing Your Own and

Initially during XYZ, I failed to truly acknowledge the their reflective journals. I would skim over them, find a simple area of interest and write a few comments on it. CT SP SH **I felt** as if it was hard to connect and empathise as my peer was involved in Electrical Engineering and **I** did not entirely **understand** the jargon of what he was **writing**, not could I acknowledge whether a task he was completing was very comprehensive or not. SH However once **I became** quite settled into my role and could see from a more external angle, I found his notebook and network as very helpful during my time. As I started my internship a week after first year exams concluded, I have no friends from my classes to speak to for advise or comparisons for roles.

**Figure 1: AWA's user interface highlights sentences in the student's text which match XIP's rhetorical patterns. 'Function Keys' such as SH signal the function that the sentence appears to be playing. Mousing over the highlight displays a prompt reminding the user of the meaning of the F key.**

interpret the learning and change in the student that has taken place.

The Engelhard Project for Connecting Life and Learning at GU, which aims to increase student well-being and deepen engagement, has been using reflective writing for ten years in over 325 courses, creating a corpus of thousands of student reflections in over 28 disciplines. A sample of this corpus, 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 systematically assess the depth and extent to which reflection, and by extension learning and change triggered by the well-being module and discussion, was occurring for these students.

# 4. MODELLING REFLECTIVE WRITING

## 4.1 NLP platform: XIP

We use the Xerox Incremental Parser (XIP) [1] for automated reflective writing analysis. XIP is a linguistic analysis engine and rule writing formalism, which has provided the platform for the development of high-performance English dependency parsing of general 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, core XIP processing performs general semantic analysis functions like named entity recognition [5] and semantic normalization [4]. The maturity of the syntactic and semantic parsing capability is evidenced by its applications for a wide variety of NLP tasks including information extraction (e.g. [11]), sentiment analysis (e.g. [6]) and discourse analysis (e.g. [22]).

XIP includes a 'salient sentences' module that models and detects relevant rhetorical moves in *analytical* writing genres like scientific and scholarly research articles, and research reports [8, 12, 23]. This provides reliable dependency parsing, and an integrated set of NLP capabilities that provide the necessary resources to build patterns for capturing features of analytical writing. See [25] for a more detailed rationale for the use of the analytical writing parser in education, and a prototype dashboard, while [25] reports preliminary evaluation in the context of an analytical writing assignment. The reflective writing parser documented in this paper is an extension of this XIP module.

## 4.2 AWA: an end-user application onto XIP

This work is part of a broader development effort at UTS to rapidly prototype writing analytics of different sorts with staff and students. Six months prior development effort, in close partnership with academic staff, had created a web application called *Academic Writing Analytics (AWA)* providing an educational user interface onto XIP.[2] This enables a piece of writing to be submitted for analysis, and the raw output from the parser is rendered in AWA as interactive highlighted text (illustrated in Figure 2).

## 4.3 Related approaches

Although reflective writing has been studied widely, little work has been devoted to its automated analysis. Besides 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. 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, *et al.* [27] developed a rule-based categoriser that decides if a text is reflective or not. Based on theoretical research in reflective writing they proposed a list of five "elements of reflection": *Description of experience, Personal experience, Critical analysis, Taking perspectives into account* and *Outcome of reflective writing*. These elements are associated with a set of

[2] AWA information: http://utscic.edu.au/tools/awa

indicators, which are used in 16 rules to detect reflective sentences. E.g. 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. The input texts are categorised 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 paid annotators (via Amazon Mechanical Turk) to evaluate the presence of the reflective elements in texts. The 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 as well as the indicators overlap with the rubrics and patterns described in Section 5.2. The major difference between the two systems is the implementation framework and the evaluation method. Whereas Ullmann et al use an array of different tools for detecting the different indicators of the reflective elements, and an independent rule formalism, 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 categorisation 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.*, and this paper, Gibson *et al.* [9] focus not 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”*), plus moves which may map to the *contrast* sentence type described in this paper (further work is needed to clarify this). Their *Anomaly Recontextualization* approach thus seeks to formalise the distinctive human ability to recognise and make sense of information which is apparently anomalous, until one reframes the context. 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”.

# 5. ITERATIVE DESIGN METHODOLOGY

To summarise, at this point we had now implemented an alpha prototype application. The availability of an independently annotated corpus at Georgetown University offered the chance to conduct a systematic evaluation. We now describe a rapid prototyping methodology for formalising rubrics into executable patterns in XIP. 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. The first step in our process was for the UTS Academic Literacies Researcher (ALR) who was affiliated with the 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 ALR had consulted with one of the subject coordinators. Through a combination of prior scholarship in the field to contextualize research for practitioners [14, 19], direct analysis of engineering students’ reflective reports, and discussion with the subject coordinator, the ALR designed the rubric to identify linguistic features and textual moves commonly associated with deep or significant reflections (Table 1).

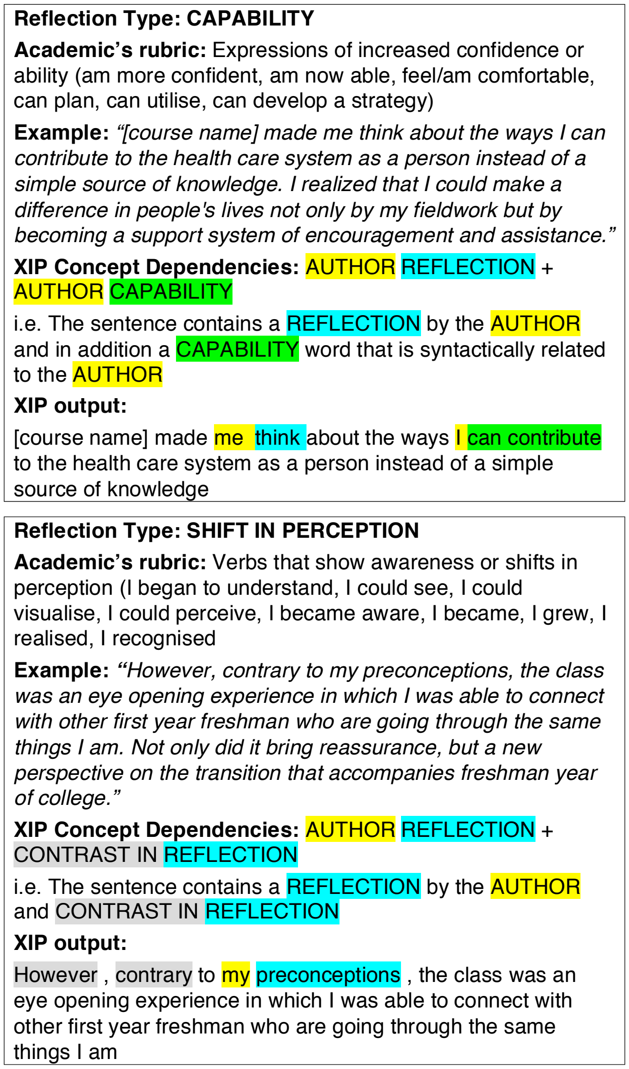
**Table 1: Rubrics for reflective writing**

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 utilise, 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 visualise, I could perceive, I became aware, I became, I grew, I realised, I recognised*
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 realised 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 next step involved modelling the rubrics as patterns, and encoding them into XIP. The patterns consist of meta-expressions, the most basic of which 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", "the suggestion that I put forward". 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. These lexicons are taken partly from the rhetorical parser previously developed, partly from the rubrics, and partly from the corpora and various synonym lists. The lexicons are evolving through the use of the AWA: as new words come up, they can be added to enlarge the coverage of the analysis. Since the parser performs dependency analysis, we could develop rules that identify the instantiations of the meta-expressions in the sentences.

Figure 2 illustrates how meta-expressions model two of the AWA reflective sentence categories using the rubrics: The category *Capability* includes the AUTHOR's REFLECTION on HER CAPABILITY, and the category *Shift in Perception* contains the AUTHOR's REFLECTION involving CONTRAST IN her REFLECTION.



**Reflection Type: CAPABILITY**

**Academic's rubric:** Expressions of increased confidence or ability (am more confident, am now able, feel/am comfortable, can plan, can utilise, can develop a strategy)

**Example:** *"[course name] made me think about the ways I can contribute to the health care system as a person instead of a simple source of knowledge. I realized that I could make a difference in people's lives not only by my fieldwork but by becoming a support system of encouragement and assistance."*

**XIP Concept Dependencies:** AUTHOR REFLECTION + AUTHOR CAPABILITY

i.e. The sentence contains a REFLECTION by the AUTHOR and in addition a CAPABILITY word that is syntactically related to the AUTHOR

**XIP output:**

[course name] made me think about the ways I can contribute to the health care system as a person instead of a simple source of knowledge

**Reflection Type: SHIFT IN PERCEPTION**

**Academic's rubric:** Verbs that show awareness or shifts in perception (I began to understand, I could see, I could visualise, I could perceive, I became aware, I became, I grew, I realised, I recognised

**Example:** *"However, contrary to my preconceptions, the class was an eye opening experience in which I was able to connect with other first year freshman who are going through the same things I am. Not only did it bring reassurance, but a new perspective on the transition that accompanies freshman year of college."*

**XIP Concept Dependencies:** AUTHOR REFLECTION + CONTRAST IN REFLECTION

i.e. The sentence contains a REFLECTION by the AUTHOR and CONTRAST IN REFLECTION

**XIP output:**

However , contrary to my preconceptions , the class was an eye opening experience in which I was able to connect with other first year freshman who are going through the same things I am

**Figure 2: From informal rubrics for good reflective writing, to formal patterns in XIP.**

As can be seen, the XIP categories use the examples in the rubrics as a basis for developing the meta-expressions. Altogether we have set up the following categories based on the rubrics: *Setting Context* (Table 1: 1st and 6th rubrics), *Specific Reflection* (2nd and 3rd rubrics), *Capability* (4th and 5th rubrics), and *Shift in Perception* (6th rubric). Any words listed in a given rubric that are not mentioned in the categories all contribute to the lexicons.

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.

## 5.3 Independent reflective writing corpus

A corpus of 30 pieces of student reflective writing (containing 382 sentences) was collected and anonymised, selected from university courses that were part of the well-being project at GU described above. Academic staff and linguistics graduate students coded each writing submission as shallow 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). Sentence-level highlighting was used to identify evidence in support of the overall code assigned. Coding consisted of trial and revision of rating rubrics, independent coding, subsequent discussion, and finally shared agreement upon coding.

An example of a shallow reflection sentence is: *"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 just described. In this sense, they were coding 'freely' as educators, rather than to test the parser.

# 6. PARSER EVALUATION

We now describe the methodology by which we evaluated the parser. As part of the iterative development design, we have tested two versions to date.

## 6.1 Results (first iteration)

The quality of classification performed by this first version of the parser was tested on the independently annotated GU corpus.

- TP (true positive) = a sentence labeled as reflective both by the parser, and the human analyst
- TN (true negative) = a sentence not labeled as reflective either by the parser or the human analyst
- FP (false positive) = a sentence labeled as reflective by the parser, but not by the human analyst
- FN (false negative) = a sentence labeled 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 an overall indicator *F1*.

| | | **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)* | |

**Table 2. Results of the first evaluation**

Considering the fact that XIP's development and the GU evaluation were entirely independent, these results were promising. We had 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".

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, discussed in more depth below under false positives, where the human annotator was coding the meaningful details that followed the reflective set-up identified by the parser. 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. The following sentence conveys other people's point of view:

> ***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 utilise, can develop a strategy)" erroneously classified the following sentence:

> *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 temporarily exclude the *Capability* type, because of time constraints for new rule development.

A major result of this first iteration was that it gave rise to introducing more subtle rules for filtering out shallow reflections from the 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*** *really* ***enjoyed*** *the freedom of being able to pick whatever science-related topic* ***interested me****.*

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 which go beyond expressing personal views about a context and take into account the *Viewpoints* of other stakeholders (see Figure 2 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: accuracy did not decrease significantly, which is promising since the new evaluation corpus had almost ten times as many sentences as the first, which increases the number of potential new words that might not have been recognised by XIP. As for the degradations in other indices, we discuss this in Sec. 7.2.

| | | **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)* | |

**Table 3. Results of second test. Brackets show the improvement with respect to first iteration results in Table 2.**

| | | **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)* | |

**Table 4. Test results on a larger corpus. Brackets show the degradation with respect to Table 3.**

### 6.3 Classifying shallow reflections

We have two indirect indications that the parser could detect shallow reflections. Firstly, we compared the ratio of sentences labelled as shallow reflection 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 reflection 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 reflection, with the remaining 160 uncoded (i.e. by implication, shallow). Although more rigorous evaluation is necessary, these two tests are indicative that the shallow reflection classifier may add value to the analysis.

## 7. DISCUSSION AND FUTURE WORK

### 7.1 Incremental rollout strategy

The first step in validation has been to build the confidence of reflective writing experts that the XIP parser has a classification scheme based in sound pedagogy and scholarship. The second step has been to quantify the performance quality, and we are encouraged that the parser, developed from scholarship in engineering reflection, is able to produce coherent results on a test corpus from other disciplines.

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. Based on other testbeds currently under way, the approach will most likely be a combination of private experimentation by students, survey and interview data, and detailed user experience evaluations using video recording and think-aloud protocols.

### 7.2 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 the light of feedback from the Georgetown University academics who had performed the hand coding. As the academics explained when they were commenting 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 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.

Another FP pattern was where 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.

A third pattern was where 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 picking up on 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 that are being made in that piece.

## 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 student is thus only fooling themselves, and in other contexts when we give AWA briefings to students, we emphasise 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 should be taught to students as guidelines, 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, and each other, 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.

## 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 UTS Academic Literacies Researcher (ALR), and the academics in UTS and Georgetown University, 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 ALR with expertise in reflective writing 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 Georgetown University 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 changes being made.

The user interface went through rapid prototyping with the ALR (and many other UTS academics testing it for their texts), using think-aloud walkthroughs. The resulting design served as a sufficiently intuitive rendering that the Georgetown University team had no difficulty in understanding how to make sense of it when reviewing and critiquing output.

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 (i) to give direct feedback to AWA on the usefulness of the sentences it is highlighting, and (ii) 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 [7, 13].

In principle this 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.

# 8. CONCLUSIONS

We have introduced the distinctive features and purposes of reflective writing as practiced 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. Recognising and understanding this evidence base sets the context for any learning analytics design efforts.

Given the challenges of teaching, learning and assessing academic reflective writing, we have identified the potential that NLP combined with a good user experience can play. 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 educators engaged with the AWA team do not feel threatened by this kind of machine intelligence; delivered in this form they see its potential to address weaknesses in the current system. 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. This feedback could also form the basis for discussion with peers and/or tutors, a provocation for sharing their understandings of what deep reflective writing 'looks like' — especially now that it can be made visible in new ways.

## 9. REFERENCES

[1] Aït-Mokhtar, S., Chanod, J-P., and Roux, C. 2002. Robustness beyond shallowness: incremental dependency parsing. *Natural Language Engineering*, 8, 2/3 (Aug. 2002), 121-144. DOI= http://dx.doi.org/10.1017/S1351324902002887

[2] Boud, D., Keogh, R., and Walker, D. 2013. *Reflection: Turning Experience into Learning*. Routledge, Abingdon, Oxon.

[3] Boud, D. and Walker, D. 1998. Promoting reflection in professional courses: the challenge of context. *Studies in Higher Education*, 23, 2 (Aug. 2006), 191-206. DOI= http://dx.doi.org/10.1080/03075079812331380384

[4] Brun, C., and Hagège, C. 2003. Normalization and paraphrasing using symbolic methods. *Proceedings of the Second International Workshop on Paraphrasing*, Volume 16. Association for Computational Linguistics, 2003. DOI= http://dx.doi.org/10.3115/1118984.1118990

[5] Brun, C., and Hagège, C. 2004. Intertwining deep syntactic processing and named entity detection. *Advances in Natural Language Processing*. Springer: Berlin, 2004. 195-206. DOI= http://dx.doi.org/10.1007/978-3-540-30228-5_18

[6] Brun, C., Popa, D.N., and Roux, C. 2014. XRCE: Hybrid classification for aspect-based sentiment analysis. *Proc. 8th International Workshop on Semantic Evaluation (SemEval 2014)*, 838–842. (Dublin, Ireland, August 23-24, 2014).

[7] Burnett, M.M., and Scaffidi, C. 2011. End-User Development. *The Encyclopedia of Human-Computer Interaction, 2nd Edition.* (Eds. Mads Soegaard and Rikke Friis Dam). Interaction Design Foundation: https://www.interaction-design.org/literature/book/the-encyclopedia-of-human-computer-interaction-2nd-ed

[8] De Liddo, A., Sándor. Á., and Buckingham Shum, S. 2012. Contested collective intelligence: Rationale, technologies, and a human-machine annotation study. *Computer Supported Cooperative Work (CSCW*) 21.4-5 (2012): 417-448. DOI= http://dx.doi.org/10.1007/s10606-011-9155-x

[9] Gibson, A. and Kitto, K. 2015. Analysing reflective text for learning analytics: an approach using anomaly recontextualisation. *Proceedings of the Fifth International Conference on Learning Analytics & Knowledge*. ACM: NY. DOI= http://dx.doi.org/10.1145/2723576.2723635

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

[11] Huang, Z., ten Teije, A., van Harmelen, F., and Aït-Mokhtar, S. 2014. Semantic representation of evidence-based clinical guidelines. *Proc. 6th International Workshop on Knowledge Representation for Health Care*. 2014: Vienna (21 July 2014). p. 78-94. DOI= http://doi.org/10.1007/978-3-319-13281-5_6

[12] Lisacek, F., Chichester, C., Kaplan, A., and Sandor, Á. 2005. Discovering paradigm shift patterns in biomedical abstracts: application to neurodegenerative diseases. In *Proceedings of the First International Symposium on Semantic Mining in Biomedicine* (SMBM) (41-50).

[13] MacLean, M., Kathleen Carter, Lövstrand, L., and Moran, T. 1990. User-tailorable systems: pressing the issues with buttons. *Proc. Conference on Human Factors in Computing Systems*. ACM, New York, 175-182. DOI= http://dx.doi.org/10.1145/97243.97271

[14] Moon, J. 2010. *Reflective Learning Workshop* (Handout 10/07), University of Worcester, UK. http://worc.ac.uk /edu/documents/Jenny_Moon_RefLearnlong07.doc

[15] Reidsema, C., Goldsmith, R., & Mort, P. 2010. Writing to learn: reflective practice in engineering design, *Proceedings of ASEE Symposium Singapore*, October 18-21

[16] Rodgers, C. 2002. Defining reflection: Another look at John Dewey and reflective thinking. *Teachers College Record.* 104.4 (2002): 842-866. http://www.tcrecord.org/content.asp?contentid=10890

[17] Russell, R. 2001. Reflection in higher education: A concept analysis. *Innovative Higher Education.* 21.6 (2001). DOI= http://dx.doi.org/10.1023/A:1010986404527

[18] Ryan, M. 2011. Improving reflective writing in higher education: a social semiotic perspective, *Teaching in Higher Education*, 16, 1 (Jan. 2011), 99-111. DOI= http://doi.org/10.1080/13562517.2010.507311

[19] Ryan, M. 2010. The 4 Rs 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

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

[21] Sándor, Á., Kaplan, A., and Rondeau, R. 2006. Discourse and citation analysis with concept-matching. *International Symposium: Discourse and Document*. (Caen, 15-17 June 2006).

[22] Sándor, Á. 2007. Modeling metadiscourse conveying the author's rhetorical strategy in biomedical research abstracts. *Revue Française de Linguistique Appliquée,* 12, 2, 97-108.

[23] Sándor, Á., and 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 (Singapore, 2-7 Aug, 2009).

[24] 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. DOI= http://dx.doi.org/10.1023/A:1003196224280

[25] Simsek, D., Buckingham Shum, S., Sándor, Á., Liddo, A. D., and Ferguson, R. 2013. XIP Dashboard: Visual analytics from automated rhetorical parsing of scientific metadiscourse. *1st International Workshop on Discourse-Centric Learning Analytics, 3rd International Conference on Learning Analytics & Knowledge* (Leuven, BE, Apr. 8-12, 2013).

[26] Sumsion, J. and Fleet, A. 1996. Reflection: can we assess it? Should we assess it?. *Assessment & Evaluation in Higher*

*Education*, 21, 2 (July. 2006), 121-130. DOI= http://dx.doi.org/10.1080/0260293960210202

[27] Ullmann, T. D., Wild, F. and Scott, P. 2012. Comparing automatically detected reflective texts with human judgements. In *2nd Workshop on Awareness and Reflection in Technology-Enhanced Learning*, Saarbruecken, Germany, September 2012, 101-116.

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

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