Reflective Writing Analytics - Empirically Determined Keywords of Written Reflection

Thomas Daniel Ullmann
Institute of Educational Technology
The Open University
MK7 6AA
+44 (0)1908 655482
t.ullmann@open.ac.uk

ABSTRACT

Despite their importance for educational practice, reflective writings are still manually analysed and assessed, posing a constraint on the use of this educational technique. Recently, research started to investigate automated approaches for analysing reflective writing. Foundational to many automated approaches is the knowledge of words that are important for the genre. This research presents keywords that are specific to several categories of a reflective writing model. These keywords have been derived from eight datasets, which contain several thousand instances using the log-likelihood method. Both performance measures, the accuracy and the Cohen's κ, for these keywords were estimated with ten-fold cross validation. The results reached an accuracy of 0.78 on average for all eight categories and a fair to good inter-rater reliability for most categories even though it did not make use of any sophisticated rule-based mechanisms or machine learning approaches. This research contributes to the development of automated reflective writing analytics that are based on data-driven empirical foundations.

CCS Concepts

• Information systems~Content analysis and feature selection • Computing methodologies~Natural language processing • Computing methodologies~Discourse, dialogue and pragmatics • Computing methodologies~Cognitive science • Computing methodologies~Supervised learning • Computing methodologies~Machine learning approaches • Computing methodologies~Feature selection • Computing methodologies~Cross-validation • Applied computing~Distance learning • Applied computing~E-learning

Keywords

Reflective writing; natural language processing; reflective writing analytics; automated detection of reflection.

1. INTRODUCTION

Reflective writing [12] is an omnipresent educational practice. It is part of the teaching curricula of many countries and disciplines [5]. The analysis and the assessment of reflective writings has been the focus of many studies. Early research indicated problems with the reliability of the assessment of reflective writings. Resent research, however, showed many cases that indicated the possibility of reliable content analysis of reflective writings [for an overview, see 14, 19].

Despite this success, the manual analysis of reflective writing remains a labour intensive and costly process, constraining the offering of such learning practices. With the advancements of computers, automated methods to analyse writings promise fast, large at scale, and reliable identification of educational constructs. Influential research areas are, for example, automated essay assessment [4] and discourse analysis [6]. Although it is widely acknowledged that reflective writing is of importance for educational practice, there has not been much research studying automated methods to analyse reflective writings. Only recently has research begun to investigate methods to automate the detection of reflection in texts.

This study contributes towards the research of automated analysis of reflection in texts by investigating the quality of empirically derived keywords that are indicative of reflection expressed in texts. Such keywords are highly important as they often form the nucleus of automated text analysis systems. This study is based on a comprehensive model of reflective writing [19, 21]. The model consists of frequently used categories to analyse reflective writing. The method generated a set of keywords for each category of the model using a data-driven approach [16]. We determined the reliability of these keywords for each category using a cross-validation approach. These keywords are a useful building block for reflective writing specific dictionaries, which can be programmatically used to gauge the indication of reflection.

This study builds on previous research of the author [20] and extends it based on the number of investigated model categories and the proposed method to estimate the performance of these keyword sets.

2. AUTOMATED APPROACHES

The general aim of research about reflection detection is the study and the development of methods that can be used to automatically identify important aspects of reflection in texts. Three reflection detection approaches have been identified in the literature [19], the dictionary-based, the rule-based, and the machine learning based approach. The focus of this research is on the dictionary-based approach. Broadly speaking, the dictionary-based approach uses defined lists of words or groups of words (the dictionaries) to automatically count the frequency of dictionary word occurrences in texts. Each dictionary resembles a category of the research object in question. The raw frequencies or metrics derived

LAK '17, March 13-17, 2017, Vancouver, BC, Canada

DOI: http://dx.doi.org/10.1145/3027385.3027394

therefrom are often used as an indicator for the model category. Often, experts determine which words or word groups belong to a dictionary. The rule-based approach uses manually constructed rules and a rule engine to infer knowledge about the text. Often, dictionaries are used in combination with the rule-based system. There is some research regarding rule-based methods to detect reflection [7, 17, 21]; however, it is less researched than the dictionary-based approach. The last and least researched approach uses machine learning algorithms to detect reflection from texts [19].

2.1 Dictionary-based Approaches

Most research has applied the dictionary-based approach to gauge insights about reflection from texts, and all the following studies used this approach but differed with regards to their conceptualisation of reflection. In an educational context, Bruno et al. [2] investigated 'mental' acts in the context of reflective practice. They defined several dictionaries with experts to study the frequency of the categories 'cognition', 'emotion', and 'volition' in student journals. Chang et al. [3] focused on four categories, namely 'cognition', 'memory', 'emotion', and 'evaluation'. They found that 'cognition' is the most frequently found dictionary type in reflection journals. The system consisting of several dictionaries compiled by Ullmann [18, 21] marked texts as reflective or non-reflective (descriptive). The evaluation showed that the human coders rated the texts similarly to the annotations of the system. Kann and Högfeldt [8] used several dictionaries for a longitudinal study of reflective writings. Lin et al. [10] leveraged the Linguistic Inquiry and Word Count (LIWC) tool, which is a system built on top of several dictionaries. They used it to research how different genders used specific words in reflective writings. Research conducted in a psychological context investigated patterns of key moments of psychoanalytical sessions [11]. One of these patterns was described as the 'reflective pattern'.

All these approaches use defined dictionaries or newly created dictionaries to analyse texts automatically. Similarly, there are methods used in linguistic research that categorise evidence found in texts according to reflection categories [for example, see 1]. Their difference to dictionary-based approaches is that in most cases the texts are manually annotated per certain categories. Despite that, this research generates knowledge about the association of linguistic features and reflection categories.

2.2 Keyword Detection

The research outlined above uses experts to manually define the words that represent the dictionaries. The approach used here is different to this 'expert' approach, as it selects words based on an empirical method and not based on expert judgements. The approach is based on the comparison of the frequency of words occurring in two datasets. A word has 'keyness' if it is frequent in one dataset but not in the other. These words have been described as 'keywords' [9]. There are several approaches to determining keywords [9]. Here, we use the log-likelihood approach described in Rayson [16].

Ullmann [20] used this approach in the context of reflective writing. The calculated keywords were based on a dataset that contained highly agreed reflective and descriptive sentences, and the evaluation reported the words with the highest 'keyness' for reflection. The interpretation of these keywords within a model of reflective writing and selected example sentences of keywords in their context corroborated the face validity of the approach. This study goes beyond the research of Ullmann [20] insofar as it determines the keywords of several categories related to reflective writing and not only one. Furthermore, this research provides an empirical estimate of the performance of these dictionaries of keywords.

3. REFLECTION DETECTION MODEL

This study uses the model for reflection detection [19] as its theoretical foundation. This model was derived from 24 models that have been studied in the context of the analysis of reflection in writings. The constituents of these models were analysed and categorised per their commonalities. The result of this synthesis is a model consisting of seven categories (eight when counting both outcome sub-categories), and many of the models used to analyse reflective writings had these categories in common. The model consists of the following seven categories. The following high-level descriptions of these categories stem from Ullmann's model [19].

Reflection: Many of the models used to analyse reflective writings described levels of reflection. Their common denominator are two levels ranging from the lowest level of reflection often called a descriptive writing—a writing that showed no presence of reflection to the highest level—a deeply reflective writing [19]. **Description of an experience**: 'This category captures the subject matter of the reflective writing' [19]. **Feelings**: 'Often, the feeling of being concerned, having doubts, feeling uncertain about something, or frustration are reasons for a reflective thought process. However, feelings such as surprise or excitement are also mentioned' [19]. **Personal beliefs**: 'Reflection is often from a personal nature. This is about one's assumptions, beliefs, the development of a personal perspective, and the knowledge of self' [19]. **Recognizing difficulties**: 'Expressing an alert, critical mindset is an important part of reflective writing. A critical stance involves being aware of problems and being able to identify or diagnose such problems' [19]. **Perspective**: 'The writer considers other perspectives. For example, the perspective of someone else, theory, the social, historical, ethical, moral, or political context' [19]. **Outcome - lessons learned and future intentions**: 'Retrospective outcomes were: Descriptions of the lessons learned, better understanding of the situation or context, new insights, a change of perspective or behaviour, and awareness about one's way of thinking. Prospective outcomes were: An intention to do something, and planning for the future' [19].

The aim of this research is to identify keywords for these categories and to evaluate their reliability.

4. METHOD

4.1 Datasets

The datasets for each category are based on previous research from the author [19]. These datasets are mainly constructed from a subset of the British Academic Written English Corpus (BAWE), which is a corpus of academic student writings [13]. The BAWE corpus was selected as it is publicly available for research, it contains samples of typical academic student writings from many university disciplines, and it contains examples of reflective writings.

Each sentence of the sample was coded by eight coders on average per the categories of the model. A sentence was included into the dataset only if a four-fifths majority of coders agreed that the sentence represented the category (class 1) or that the sentence did not represent the category (class 2). Ullmann [19] described the datasets as reliable and valid.

Table 1 shows the size of all datasets and their respective class distribution. For example, the dataset from the 'Reflection' category consists of 2347 sentences, which 603 sentences are highly agreed as being reflective (class 1) and 1744 sentences are highly agreed as being non-reflective/descriptive (class 2).

Table 1. Size of datasets and class distribution

Dataset	N	Class 1	Class 2
Reflection	2347	603	1744
Experience	3392	1563	1829
Feeling	2672	811	1861
Belief	2303	1188	1115
Difficulty	2717	1392	1325
Perspective	2028	330	1698
Learning	1882	699	1183
Intention	3755	347	3408

4.2 Performance Estimates of Keywords

The performance of the derived keywords from the datasets was estimated using 10-fold cross-validation. The dataset was split randomly into 10 equally sized sets from which we generated three sets, the training set (80% of the data), the validation set (10%), and the test set (10%). The eight parts of the training set were used to determine the 'keyness' of each word. For this, we calculated the log-likelihood of each word, and ordered the words from the largest log-likelihood ratio to the smallest, including only words that got used more often in sentences of class 1. This ordered list was used to determine the best candidate keywords based on the validation data set. Starting with the first keyword of the previously generated list, the instances of the validation set were classified as class 1 if they contained the keyword; otherwise, they were labelled as class 2. The comparison of the classified sentences with the class labels generated by the human coders served as inter-rater reliability estimates for this keyword. Subsequently, the keyword with the highest and second highest log-likelihood ratio served to determine if the sentences belonged to the class and to calculate the inter-rater reliability. This process continued for the remaining keywords. The set of keywords that had the highest inter-rater reliability measured with Cohen's κ was chosen as the candidates to be tested on the novel data of the remaining test set. The last part, the test set, was used to determine the performance of the previously determined set of candidate keywords on novel data.

This process of calculating the log-likelihood of all words on eight parts of the dataset, finding the best candidate keywords on the validation dataset, and estimating the performance of these keywords on the test set was repeated 10 times. The result section shows the mean and the standard deviation of these 10 repetitions. As measurements, we chose the accuracy (per cent agreement) and Cohen's κ. Both have been frequently used in similar research contexts [19].

4.3 Statistical Software

The R project for statistical computing [15] served for all calculations. The R tm package was used to pre-process all instances to the lower case and to tokenize the sentences to unigrams. We calculated Cohen's κ and the percent agreement with the R scripts provided by Gwet[1]. We developed our own function to calculate the log-likelihood ratio according to the information found in Rayson [16] and all other scripts needed to perform the calculations.

[1]http://www.agreestat.com/r_functions.html

5. RESULTS

Table 2 shows the aggregated performance measured with Cohen's κ and the accuracy of the set of keywords estimated on the test folds for each category. Nfold shows the average amount of instances of the 10 test folds.

Table 2. Accuracy and Cohens's κ for each category.

	Nfold	Accuracy		Cohen's κ	
Category	**Mean**	**Mean**	**SD**	**Mean**	**SD**
Reflection	234.7	0.83	0.03	0.59	0.07
Experience	339.2	0.82	0.02	0.65	0.03
Feeling	267.2	0.79	0.03	0.56	0.05
Belief	230.3	0.70	0.03	0.39	0.06
Difficulty	271.7	0.73	0.03	0.47	0.05
Perspective	202.8	0.74	0.06	0.28	0.04
Learning	188.2	0.67	0.05	0.34	0.06
Intention	375.5	0.93	0.01	0.51	0.10
Average	263.70	0.78	0.03	0.47	0.06

The accuracy (often also called the percent agreement) ranges from 67% to 93%. All categories except 'Perspective' were above the baseline accuracy. The baseline accuracy is the accuracy of a method that always predicts the majority class as true. Table 1 provides all the values needed to calculate the baseline accuracy, which are the following: 'Reflection' 0.74, 'Experience' 0.54, 'Feeling' 0.70, 'Belief' 0.52, 'Difficulty' 0.52, 'Perspective' 0.84, 'Learning' 0.63, and 'Intention' 0.91. Based on these baseline values, only the keywords derived from the 'Perspective' dataset had a lower accuracy than the baseline. Table 2 also show that Cohen's κ ranged from 0.28 to 0.65. On the benchmark of Landis and Koch, 'Experience' reached substantial, 'Reflection', 'Feeling', 'Intention', and 'Difficulty' moderate, and 'Learning', 'Perspective', and 'Belief' fair inter-rater reliability. Per the benchmark of Fleiss, the κ values for five categories, 'Reflection', 'Experience', 'Feeling', 'Difficulty', and 'Intention', had fair to good agreement, while the other three categories performed poorly.

As outlined above, we used the validation set to determine the set of keywords with the highest log-likelihood. These are the candidate keywords that were used for the assessment of the test data performance, as they will likely perform on the test dataset similarly to the validation dataset. This process was repeated ten times. Therefore, for each iteration of the cross-validation set, ten sets of keywords have been generated for each category. Often, these keywords are the same from iteration to iteration, but they are sometimes different. Instead of reporting all ten sets of keywords for each category, we calculated the percentages of keywords appearing in all folds. For example, a keyword that appeared in eight out of ten folds has a percentage of 80%. We only show keywords that were used at least in 50% of the folds.

Reflection: The two keywords were 'I' (100%) and 'me' (60%). 'I' had the highest log-likelihood and the highest Cohen's κ on four folds, while the combination of 'I' and 'me' yielded, in six instances, the highest κ on the test sets.

Experience: The derived experience keywords were the singular first-person pronouns 'I' (100%) and 'me', (100%) in addition to

the plural first-person pronoun 'we' (90%). Furthermore, the past tense auxiliary verbs 'was' (100%), 'had' (100%), 'were' (90%), and 'did' (50%) were present.

Feeling: Keywords were the singular first-person pronouns 'I' (100%) and 'me' (60%) and the two verbs 'feel' (80%) and 'felt' (60%) expressing feelings.

Belief: The best candidate keywords derived from the 'Belief' dataset were the first-person pronouns 'I' (100%), 'my' (80%), and 'it' (50%); the sensing and thinking verbs 'feel' (100%), 'believe' (90%), and 'think' (80%); as well as the auxiliary verb 'have' (60%).

Difficulty: The keywords generated from the 'Difficulty' dataset have been manifold. They were the conjunctions 'because' (100%), 'but' (100%), 'if' (90%), and 'although' (70%); the nouns 'lack' (90%), 'problems' (80%), and 'situation' (80%); the adjectives 'difficult' (100%), 'due' (100%), and 'wrong' (80%); the verbs 'trying' (60%), 'felt' (50%), and 'made' (90%); the auxiliary verbs 'did' (100%), 'didn't' (100%), 'don't' (60%), 'have' (100%), 'could' (100%), 'would' (100%), and 'may' (100%); the adverbs 'still' (70%), 'not' (100%), and 'however' (100%); and the third-person pronoun 'it' (60%).

Perspective: The keywords generated from the 'Perspective' dataset were the third person pronouns 'they' (100%), 'she' (50%), and 'his' (50%); the verbs 'felt' (90%), 'said' (80%), and 'understand' (50%); the auxiliary verbs 'may' (100%), 'might' (50%), and 'would' (50%); the adjective 'aware' (50%); and the conjunction 'that' (80%).

Learning: The generated keywords for 'Learning' were the first-person pronouns 'me' (100%) and 'I' (70%); the nouns 'future' (100%), and 'experience' (90%); the verbs 'learnt' (100%) and 'have' (80%); and 'better' (100%), which was used as either adjective or adverb.

'Intention' had only one keyword, the modal verb 'will' (100%).

Overall, the amount of keywords per category varied, but mostly a combination of few keywords had the highest performance.

6. DISCUSSION

The evaluation of this study showed that the chosen method generates word lists from the datasets that detect categories of reflective writing with fair to good reliability for most categories of the reflective writing model. Most categories had an accuracy above the baseline outperforming chance agreement. Both measurements showed that keywords can detect model categories of reflective writing.

Some of the keywords of categories, such as 'Experience', 'Reflection', and 'Feeling', performed better than others, for example, 'Perspective' and 'Learning'. This suggests that within the context of the current set-up of the experiment, some categories of reflective writing are harder to automate than others using keywords as dictionaries.

The evaluation of the keywords showed that with a relatively small set of keywords, we can detect categories of reflective writing with fair to good reliability. Although these keywords have been determined by an algorithm and not by experienced humans, they intuitively make sense in the context of their categories. They may not be representative of the whole category, but they represent the concept to a degree. This representativeness is an indicator for face validity. I will elaborate now on this point for each category.

The two first-person pronouns 'I' and 'me' are highly indicative of sentences that have been judged as reflective compared to non-reflective/descriptive sentences. This finding is congruent with the research of Birney [1] that highlighted the importance of the 'first person voice' for reflection.

The keywords for the category 'Experience' were also the two first-person pronouns indicating that sentences describing an experience are often told from the first-person perspective. An experience is often described as an event from the past, which is supported by the unusually frequent use of the past tense verbs 'was', 'were', 'had', and 'did'.

Intuitively, it makes sense that keywords 'feel' and 'felt' together with 'I' and 'me' are used to express personal feelings.

Keywords that had an unusually high frequency in the 'Belief' dataset were again first-person pronouns as well as the cognitive and sensing verbs, such as 'think', 'believe', and 'feel'.

The noun keywords from the 'Difficulty' dataset indicate that sentences expressing difficulties discuss a 'problem', a 'lack' of something, and a 'situation'. The adjectives 'difficult' and 'wrong' specify something negative, and further negations are expressed in the keywords 'not', didn't' and 'don't'. The conjunctions 'but' and 'although' often express a contrast. 'Because' can be used to indicate causes or reasons, and 'if' can be interpreted as a sign that a writer is thinking of one or several conditions and their likely outcome. Similarly, 'would' can be used to express imagined situations, and the words 'could' and 'may' can be used to express an alternative or possibility. Lastly, 'trying' expresses an attempt to do something, which indicates that the writer is not yet in the position he or she wants to be in, which is well aligned with the description of difficulties and problems.

The keywords with the highest log-likelihood ratio calculated from the 'Perspective' dataset had many third-person pronouns, indicating that the writer was frequently referring to someone else. This someone could have 'said' something that added another perspective to the writing. Furthermore, a (new) 'understand'(ing) or 'aware'(ness) of a situation can be expressions of another perspective. The words 'may', and 'might' can indicate an expression of an alternative or possibility. Similarly, the use of 'would' can signpost what someone would do, for example, by imagining another perspective.

The 'Learning' keywords were the personal pronouns 'I', and 'me' showing a personal stance when reporting 'learning'. The verbs 'learnt' and 'have' indicate that something was learnt or that someone has done something. When reporting learning, the writer talked about an 'experience' and what was learnt for the 'future'. The writer might indicate that he or she is now 'better' in doing something.

The most frequent keyword of the 'Intention' dataset with the highest Cohen's κ was the modal verb 'will', signposting something in the future.

These examples of the automatically generated keywords in the context of their categories of a reflective writing model showed that many of the keywords are indeed relatable to their category. They might not represent the category in its entirety as one can find other words that can be plausibly related to the category. They are, however, the keywords that had the highest Cohen's κ, which was determined during cross validation. These characteristics make them important elements for tools to automatically generate indicators of reflective writing.

The results of this research must be seen in their context. The datasets are based on a text collection of academic student writings. Most of the research regarding reflective writing is based on such writings, which makes it a suitable data source. Other forms of reflective writing do exist, such as blogging, which was not investigated. Furthermore, the dataset contained only English texts. It would be interesting to investigate the reflection in different languages. This could lead to a better understanding of how reflection is expressed all over the world. The work of Kann and Högfeldt [8] and Lin et al. [10] points in this direction.

The keywords that have been identified in this research do not only provide insights for tool developers that wish to make use of dictionaries to analyse reflective writings. They can be also useful for research on rule-based and machine learning based methods for reflection detection. Rule-based approaches often combine dictionaries with rules [7, 17, 21]. These dictionaries are a core building block of such rule-based systems. Supervised machine-learning based approaches to classify texts often must manage a large set of features that are used to train the machine learning models. Dictionaries have been used to reduce the feature space to a manageable size for the training of these models.

7. CONCLUSION

Reflective writing is such an important educational practice that limitations posed by the bottleneck of available teaching time should be challenged. Several researchers see automated methods to analyse and assess reflective writing as a major step towards the solution to this problem [7, 17, 19]. Studying such methods can help explain which parts of the analysis of reflective writings can be automatised. Currently, research in automated detection of reflection is in its infancy. This research focussed on the quality with which automated methods can detect reflection in writings. This choice is underlined by the belief that it is important for any future study of the interaction between learners and automated tools to have methods in place that can reliably detect reflection from texts. This research added a missing link to this challenge, showing how data driven approach can be used to better understand reflective writing.

8. REFERENCES

[1] Birney, R. 2012. *Reflective Writing: Quantitative Assessment and Identification of Linguistic Features*. Waterford Institute of Technology.

[2] Bruno, A., Galuppo, L. and Gilardi, S. 2011. Evaluating the reflexive practices in a learning experience. *European Journal of Psychology of Education*. 26, (May 2011), 527–543.

[3] Chang, C.-C., Chen, C.-C. and Chen, Y.-H. 2012. Reflective behaviors under a web-based portfolio assessment environment for high school students in a computer course. *Computers & Education*. 58, 1 (Jan. 2012), 459–469.

[4] Dikli, S. 2006. An Overview of Automated Scoring of Essays. *The Journal of Technology, Learning and Assessment*. 5, 1 (Aug. 2006).

[5] Dyment, J.E. and O'Connell, T.S. 2010. The Quality of Reflection in Student Journals: A Review of Limiting and Enabling Factors. *Innovative Higher Education*. 35, (Mar. 2010), 233–244.

[6] Ferguson, R. and Shum, S.B. 2011. Learning analytics to identify exploratory dialogue within synchronous text chat. *Proceedings of the 1st International Conference on Learning Analytics and Knowledge* (2011), 99–103.

[7] Gibson, A., Kitto, K. and Bruza, P. 2016. Towards the Discovery of Learner Metacognition From Reflective Writing. *Journal of Learning Analytics*. 3, 2 (Sep. 2016), 22–36.

[8] Kann, V. and Högfeldt, A.-K. 2016. Effects of a Program Integrating Course for Students of Computer Science and Engineering. *Proceedings of the 47th ACM Technical Symposium on Computing Science Education* (New York, NY, USA, 2016), 510–515.

[9] Lijffijt, J., Nevalainen, T., Säily, T., Papapetrou, P., Puolamäki, K. and Mannila, H. 2014. Significance testing of word frequencies in corpora. *Digital Scholarship in the Humanities*. (Dec. 2014).

[10] Lin, C.-W., Lin, M.-J., Wen, C.-C. and Chu, S.-Y. 2016. A word-count approach to analyze linguistic patterns in the reflective writings of medical students. *Medical Education Online*. 21, (Feb. 2016).

[11] Mergenthaler, E. 1996. Emotion–abstraction patterns in verbatim protocols: A new way of describing psychotherapeutic processes. *Journal of Consulting and Clinical Psychology*. 64, 6 (Dec. 1996), 1306–1315.

[12] Moon, J.A. 2006. *Learning Journals: A Handbook for Reflective Practice and Professional Development*. Routledge.

[13] Nesi, H. and Gardner, S. 2012. *Genres across the disciplines: Student writing in higher education*. Cambridge University Press.

[14] Poldner, E., Simons, P.R.J., Wijngaards, G. and van der Schaaf, M.F. 2012. Quantitative content analysis procedures to analyse students' reflective essays: A methodological review of psychometric and edumetric aspects. *Educational Research Review*. 7, 1 (2012), 19–37.

[15] R Core Team 2014. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing.

[16] Rayson, P. 2008. From key words to key semantic domains. *International Journal of Corpus Linguistics*. 13, 4 (2008), 519–549.

[17] Shum, S.B., Sándor, Á., Goldsmith, R., Wang, X., Bass, R. and McWilliams, M. 2016. Reflecting on Reflective Writing Analytics: Assessment Challenges and Iterative Evaluation of a Prototype Tool. *Proceedings of the Sixth International Conference on Learning Analytics & Knowledge* (New York, NY, USA, 2016), 213–222.

[18] Ullmann, T.D. 2011. An Architecture for the Automated Detection of Textual Indicators of Reflection. *Proceedings of the 1st European Workshop on Awareness and Reflection in Learning Networks* (Palermo, Italy, 2011), 138–151.

[19] Ullmann, T.D. 2015. *Automated detection of reflection in texts. A machine learning based approach*. The Open University.

[20] Ullmann, T.D. 2015. Keywords of written reflection - a comparison between reflective and descriptive datasets. *Proceedings of the 5th Workshop on Awareness and Reflection in Technology Enhanced Learning* (Toledo, Spain, Sep. 2015), 83–96.

[21] Ullmann, T.D., Wild, F. and Scott, P. 2012. Comparing Automatically Detected Reflective Texts with Human Judgements. *2nd Workshop on Awareness and Reflection in Technology-Enhanced Learning* (Saarbruecken, Germany, Sep. 2012).