**Title: Advances in the field of Automated Essay Evaluation  
Authors: Kaja Zupanc and Zoran Bosnić**

*We thank the reviewers for their valuable comments and recommendations that helped us to improve our paper. In the following, we attach the list of responses to the comments.*

**Reviewer A:**

**1.) There is a lack of formal definitions, e.g. define what an essay is, and several other concepts used in the paper.**

*Several definitions and examples were added to the paper, so the main concepts are presented more clearly:*

* *Page 1, Introduction, paragraph 1, definition of an essay: “Essays are a short literary composition on a particular theme or subject, usually in prose and generally analytic, speculative, or interpretative.”*
* *Page 3, Section 3, paragraph 1, definition of response feedback: “Feedback for each specific response provides information on quality of different aspects of writing, as partial score as well as descriptive feedback.”*
* *Page 4, Section 3.2, paragraph 1, components of text mechanics: “(e.g. spellchecking errors, capitalization errors, punctuation errors)”,*
* *Page 4, Section 3.2, paragraph 1, components of discourse elements: “(e.g. title, introductory material, thesis, main idea, supporting idea, conclusion)”.*

**2.) A test should be made on all these systems on benchmark data. But the paper is publishable already without it.**

*Unfortunately most of the systems are private (their proprietary logic is not available to the public) so we were not able to perform a test to compare them. But we included the results of the comparison of nine systems on the same data set and also included results that were obtained with the same data set on two other systems.*

*The following text was added to clarify this:*

* *Page 12, Section 4.3, paragraph 3: Unfortunately we were not able to test the rest of the systems on the same data set or use independent data set to compare all of the system, since majority of the systems are proprietary and not publicly available.*

**3.) Smaller remarks:**  
- **"field" in the title should be "Field"  
- authors header: K. Zupanc et al.  
- section titles should be in small letters  
- IN ABSTRACT: "time-consuming, labor-intensive and expensive" seems a bit too much emphasized**

*We corrected the suggested mistakes.*

------------------------------------------------------------------------------------------------------------

**Reviewer B:**

**1.) Since the field of AEE is probably not familiar to the Informatica readership, the paper should attempt to make it a bit clearer what these systems can do. Explanation of the terms such as writing construct, writing mechanics, discourse elements etc., which are generally not familiar to computer scientists, would go some way towards this objective.**

*Several definitions and examples were added to the paper, so the main concepts are presented more clearly.  
Please see response 1 to Reviewer A for details.*

**2.) I think a more thorough comparison (what types of attributes are extracted from essays for evaluation) would be valuable, but I realize this may not possible since developers of these systems may not reveal adequate information.**

*This is true, most of the systems are unfortunately proprietary so we were not able to describe them in detail. Nevertheless, to clarify types of attributes, we extended the following text to describe the systems more in detail where the information is available:*

* *Page 7, Section 4.2, e-rater, paragraph 1: The attribute classes include the following [4,9] (1) grammatical errors (e.g. subject-verb agreement errors), (2) word usage errors (e.g. their versus there), (3) errors in writing mechanics (e.g. spelling), (4) presence of essay-based discourse elements (e.g. thesis statement, main points, supporting details, and conclusions), (5) development of essay-based discourse elements, (6) style weaknesses (e.g. overly repetitious words), …*
* *Page 7, Section 4.2, Intelligent Essay Assessor, paragraph 1: IEA uses LSA to derive attributes describing … The system uses approximately 60 attributes to measure above aspects within essays: content (e.g. LSA essay semantic similarity, vector length), lexical sophistication (e.g. word maturity, word variety, and confusable words), grammar (e.g. n-gram attributes, grammatical errors, and grammar error types), mechanics (e.g. spelling, capitalization, and punctuation), style, organization, and development (e.g. sentence-sentence coherence, overall essay coherence, and topic development).*
* *Page 7, Section 4.2, IntelliMetric, paragraph 1: The content attributes evaluate the topic covered, the breadth of content, and support for advanced concepts, cohesiveness and consistency in purpose and main idea, and logic of discourse. Whereas structure attributes evaluate grammar, spelling, capitalization, sentence completeness, punctuation, syntactic variety, sentence complexity, usage, readability, and subject-verb agreement*
* *Page 9, Section 4.2, Semantic Automated Grader for Essays: The novelty of the system is a set of semantic coherence attributes measuring changes between sequential essay parts from three different perspectives: semantic distance (e.g. distance between consecutive parts of an essay, maximum distance between any two parts), central spatial tendency/dispersion, and spatial autocorrelation in semantic space.*

**3.) I also have a few minor comments:  
- While the language is generally good, there are a few issues: arised -> arisen, the most of the existing -> most of the existing, widespread is not a noun, the most of the AEE research -> most of the AEE research, the several attribute classes -> several attribute classes (there are probably  
some more supeflous the-s in the paper).**

*Thank you, we corrected the suggested mistakes.*

**4.) Why would psychologists model AEE systems? I would understand if they modeled students in order to develp AEE systems, but not the other way around.**

*We corrected the confusing description to: “… and cognitive psychologists expert opinion is considered when modeling the attributes.”*

**5.) Overview of the paper at the end of Section 1 should be by section numbers; Section 3 should probably be 4.**

*We corrected the paper overview.*

**6.) What is feedback tractability and score logic for a specific response?**

*The description was supplemented to make the content presented more clearly:*

* *Page 3, Section 3, paragraph 1: “AEE systems provide higher degree of feedback tractability and score logic for a specific response Feedback for each specific response provides information on quality of different aspects of writing, as partial score as well as descriptive feedback.”*

**7.) Off-topic advisory is mentioned twice at the beginning of Section 3.6.**

**8.) Are unorthodox essays also a challenge?**

*We corrected the mistake by changing one “off-topic” occurrence to “unorthodox” to better highlight both challenges.*

**9.) I do not understand the part about substantive knowledge when talking about SAGrader.**

*The description was corrected to make the content presented more clearly:*

* *Page 9, Section 4.2, SAGrader, paragraph 1: Then the instructor creates a rubric identifying the ``desired features'' -- key elements of knowledge (set of facts) that should be included in a good response, along with relationships among those elements -- using a semantic network (SN).*

**10.) What is the proportion of the model answer knowledge ...?**

*The description of the Markit system was corrected to make the content presented more clearly:*

* *Page 10, Section 4.2, Markit, paragraph 1: It requires comprehensive knowledge in a form of one model (exemplary) answer against which the student essays are compared.*