Automated free-text assessment: Some lessons learned

Philippe Dessus, Benoît Lemaire, Mathieu Loiseau, Sonia Mandin, Emmanuelle Villiot Leclercq, Virginie Zampa

Automated free-text assessment: some lessons learned

Philippe Dessus*

Laboratoire des sciences de l'éducation & IUFM-UJF Grenoble-1,
Université Pierre-Mendès-France,
Bât SHM, 1251, av. Centrale, BP 47,
38040 Grenoble CEDEX 9, France
E-mail: Philippe.Dessus@upmf-grenoble.fr
*Corresponding author

Benoît Lemaire

LPNC-CNRS,
Université Pierre-Mendès-France,
Bât SHM, 1251, av. Centrale, BP 47,
38040 Grenoble CEDEX 9, France
E-mail: Benoit.Lemaire@upmf-grenoble.fr

Mathieu Loiseau

Laboratoire des sciences de l'éducation & LIDILEM Grenoble-3,
Université Pierre-Mendès-France,
Bât SHM, 1251, av. Centrale, BP 47,
38040 Grenoble CEDEX 9, France
E-mail: Mathieu.Loiseau@msh-alpes.fr

Sonia Mandin

Laboratoire des sciences de l'éducation & IUFM-UJF Grenoble-1,
Université Pierre-Mendès-France,
Bât SHM, 1251, av. Centrale, BP 47,
38040 Grenoble CEDEX 9, France
E-mail: Sonia.Mandin@upmf-grenoble.fr

Emmanuelle Villiot-Leclercq

IUFM de Grenoble,
30 av. Marcelin-Berthelot,
38100 Grenoble, France
E-mail: Emmanuelle.Villiot-Leclercq@ujf-grenoble.fr

Virginie Zampa

LIDILEM,
UFR des Sciences du Langage,
Université Stendhal,
BP 25, 38040 Grenoble CEDEX 9, France
E-mail: Virginie.Zampa@gmail.com

Abstract: Most e-learning systems engage successively students in reading, writing and assessment activities. In the third phase, the teacher gives feedback on student comprehension, which is often processed a long time after the others, letting the students alone with their difficulties. Thus, there is room to devise automated assessment systems on course comprehension, based on NLP techniques such as latent semantic analysis (LSA). The aim of this paper is to present some systems devised to complete this aim, which implement LSA to model learners' comprehension and/or to compare reading material (e.g., course text) with learners' summaries about it, select reading materials and predict student processes from their summaries.

Keywords: life-long learning; e-learning; personal learning environments; PLEs; learner comprehension; feedback; free-text assessment; writing; reading; latent; semantic analysis; natural language processing.

Reference to this paper should be made as follows: Dessus, P., Lemaire, B., Loiseau, M., Mandin, S., Villiot-Leclercq, E. and Zampa, V. (2011) 'Automated free-text assessment: some lessons learned', *Int. J. Continuing Engineering Education and Life-Long Learning*, Vol. 21, Nos. 2/3, pp.140–154.

Biographical notes: Philippe Dessus is a Full Professor in Educational Sciences at Grenoble University. His research focuses on cognitive tools for teaching and learning, mainly using latent semantic analysis, and he is involved in the language technologies for life-long learning project (LTfLL), an EC-funded research project from which *Pensum* is developed.

Benoît Lemaire is an Associate Professor in Computer Science from Grenoble University. His main research goal is to design computational models of cognitive processes, like human language learning, information-seeking tasks, semantic memory, or text comprehension. He is also involved in the LTfLL project.

Mathieu Loiseau is also involved in the LTfLL project, in which he works on the development of *Pensum* as a Post-doctorate Fellow. His research interests span over technology enhanced learning (TEL), computer assisted language learning in particular, natural language processing, software engineering, language didactics and educational sciences.

Sonia Mandin is a Researcher in Educational Sciences at Grenoble University. Her researches concern the design and the implementation in TEL environments of models about cognitive processes involved in writing activities such as summarising.

Emmanuelle Villiot-Leclercq is a Researcher in Educational Sciences. Her research field is settled in the context of TEL and concerns more particularly the field of educational design and scenarisation in a distant context. Since 2009, she is a member of TECFA laboratory (University of Geneva) and a Teacher in ICT at the Teachers Training Institute of Grenoble University.

Virginie Zampa is an Associate Professor in Computer Science at *Linguistique et Didactique des Langues Etrangères et Maternelles* laboratory (LIDILEM) from Grenoble University. She works on various subjects like computer environment for human learning, especially for foreign language acquisition, natural language processing and more precisely in semantic analysis, language didactics and educational sciences.

1 Introduction

What is in common to any student learning in a life-long context (henceforth LLL)? What is worth taking into account when developing computer-based assessments (e.g., Fischer, 2001; Koper, 2004)? Previous studies in that domain emphasised the following characteristics. First, LLL students often work in a collaborative group context (e.g., to lessen loneliness, Graham and Perin, 2007). Second, full-day access to e-learning courses is preferable. Third, since students often are engaged in professional activities, there is a necessity to blend distance and presence interactions. Fourth, students may have different levels of expertise/experience.

Linn (1996) elaborated a list of capabilities for students engaged in an academic context. This list is transferable to any LLL situation and emphasises two main capabilities necessary for the protagonists of such a situation. The learner has to take an *autonomous stance* toward learning, while the teacher has to take a *scaffolded knowledge integration* stance toward instruction (i.e., trying to expand the learner's ideas, helping learners to distinguish between these ideas by reflecting, organising, connecting them into coherent perspectives).

In such LLL situations computer-based assistance can mediate learners and teachers for providing just-in-time assessment tools. Most of current e-learning systems engage students in a threefold activity: *reading* some learning materials on a given course content, *writing* some ideas (e.g., the format of which can be course notes, summaries, syntheses, etc.), and *interacting* with a teacher or tutor who can give them feedback on their comprehension. The third phase is often processed a long time after the first two and some computer-based tools may be used in the meantime in order to provide just-in-time feedback. Needless to say that these current tools mostly perform shallow analysis, which makes them vulnerable to being fooled (e.g., quizzes, multiple choice questionnaires) or on the contrary to consider erroneous answers to be correct (e.g., in the field of language learning, Blanchard et al., 2009). Both aspects of this third phase result in students often ending up alone with their understanding problems. The aim of this paper is to review some computer-based environments that let learners organise themselves around these three lines of activities, and especially the third one.

2 Theoretical approaches for automated text assessment

The automated assessment of free texts has started as a new research field (Ericsson and Haswell, 2006; Williamson et al., 2006). Researches on this field are based on the three-fold assumption:

1 if a student writes about something (e.g., free texts) this will lead to learning (Klein, 1999)

2 just-in-time feedback on this writing, even imperfect, will enable students to reflect on the content taught, and eventually to revise their writing

3 making students involved in this kind of activity leads them to self-regulate their learning.

In brief, automated feedback promotes self-regulated learning, which in turn enables to build knowledge through writing. The remainder of this section elaborates on these points.

2.1 Writing-to-learn

The writing-to-learn approach is well suited for LLL. Very roughly, this approach assumes that learners can transform their writing into knowledge (Bangert-Drowns et al., 2004). The way the writing activity can promote learning is not fully uncovered yet, but, as Langer and Applebee (1987, pp.135–136) put it:

> "1. Writing activities promote learning better than activities involving only studying or reading.
>
> 2. Different kinds of writing activities lead students to focus on different kinds of information.
>
> 3. In contrast to short-answer responses, which turn information into discrete small pieces, analytic writing promotes more complex and thoughtful inquiry but on a smaller amount of information."

This assumption lets us propose tools that support several kinds of writing activities students can carry out (e.g., notes, keywords, essays, syntheses, definitions), their production being assessed by computer-based tools.

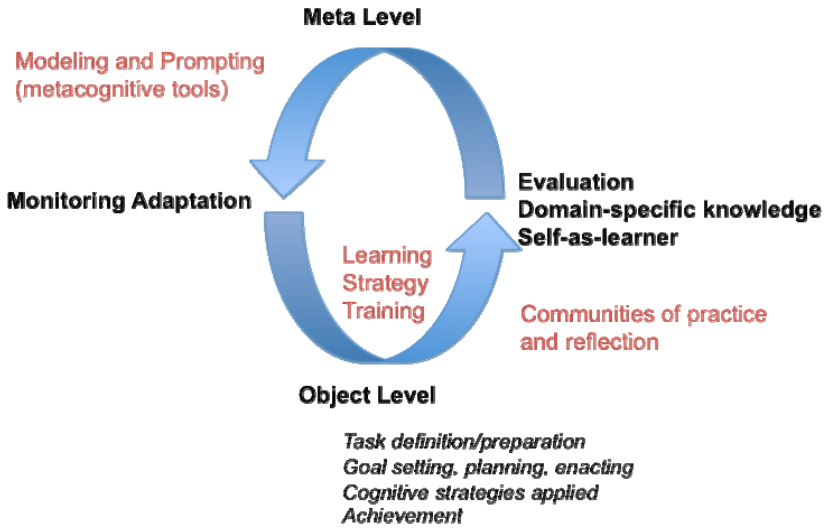
2.2 Three steps in self-regulated learning

One of the key-features of computer-based LLL is that electronic devices used by learners enable them to self-regulated their learning. As Lin (2001, p.26) stated: "The key to the success in their design [of LLL situations] was to have students experience these strategies in their own learning, explicitly compare their own performance with that of the model, and take action to revise ineffective learning approaches". Paradoxically, the literature on personal learning environments (PLEs) seldom reports models integrating self-regulated learning (SRL) processes (Vovides et al., 2007).

These authors proposed an SRL model especially dedicated to PLEs, in which students are considered to be able to perform the following activities. First, students work on the *object level*, in preparing their activity according to the ongoing task. They can also apply for some cognitive strategies for these activities to be performed. Then,

students perform a first rough assessment of their production (its adequacy, its relation to their knowledge, etc.). Third, a reflection on a meta-level allows them to perform a comparison between the latter assessment and the object level, often offered by artefacts (computer-based services, prompts, etc.), and therefore to compare their perceived level of learning with that proposed by the artefacts. Eventually, the student can perform some adaptations to their work, which in turn fuels the possible update of the object level and can be re-acted in a distance learning context.

Figure 1 Metacognitive approach to design e-learning activities (see online version for colours)



Source: Vovides et al. (2007, p.68)

3 Computer-based essay assessment

3.1 A short literature review

Computer programs for automatically assessing student free texts date back to the 60s (Page, 1966). These first systems were based on surface features like the length of the words used by the student, the number of connectives, etc. (Wresch, 1993). Surprisingly, these simple variables prove to correlate quite well with grades given by teachers. *E-rater* is such a system (Burstein et al., 1998). It scans 60 textual features, which are filled by a syntactic parser and a rhetorical pattern detector.

The next generation of systems addresses the semantic level. *Intelligent essay assessor* (Foltz et al., 1999) uses latent semantic analysis (LSA) to compare student essays to pre-graded texts. The grade given can be either the grade of the closest pre-graded essay (holistic score) or the semantic similarity to a reference essay (gold standard score). *Summary street* (Kintsch et al., 2007; Wade-Stein and Kintsch, 2004) is similar but is intended to assess summaries. Much effort has been given to design a graphical view of the feedback. *Automatic essay assessor* is a similar software (Kakkonen et al., 2004) using of part-of-speech tags to enhance feedback (Kakkonen et al., 2006).

3.2 LSA as a multipurpose free-text assessment method

Assessing automatically learners essays can be done at various levels but the most challenging is, as for NLP in general (Antoniadis, 2004), the most distant from the form, among which the semantic level. There are two kinds of semantic assessments of a learner production: *within-comparisons*, which evaluate the essay itself, and *between-comparisons*, which are intended to compare the learner production to extra materials. Both methods require a robust semantic representation of the meaning of the compared passages. Highly detailed and relevant semantic representations are very hard to achieve in spite of remarkable progress in computational linguistics and ontology processing. However, comparison can be done based on rougher representations.

Let us take an example. In order to assess a synthesis, one does not need a precise description of the meaning of a sentence such as "A lion can live a dozen years and even more in a zoo". However, the system should be able to estimate that this sentence is a good recall of the book sentence, which was "Lions live for ten to 14 years in the wild, while in captivity they can live longer than 20 years". A simple estimation of the degree of semantic relatedness between passages can thus be informative enough for essay assessment systems (provided there is a form of reference text to compare the essay with). It is worth noting that to perform consistently that type of comparison at the level of the text, one can hardly imagine a system that would not integrate semantic level comparison of some sort.

LSA (Landauer et al., 2007) is an appropriate technique for essay assessments because it provides such a measure of semantic similarity between any sequence of words. LSA analyses a large corpus of raw texts and represents each word by a high-dimensional vector (usually hundreds of dimensions) in a so-called semantic space. This representation solves the classical compositional problem, which is to represent the meaning of sequences of words that were not in the learning corpus. Any new sequence of words is indeed represented as a vector, which is just the linear combination of its words. Since everything is represented as a vector, it is straightforward to compute any semantic similarities. This is usually performed by means of the cosine function.

LSA is an interesting method for educational text assessment because its measures proved to be comparable to human judgments of similarity (Landauer et al., 2007), and Dessus (2009) listed a comprehensive overview of the use of this method in educational contexts.

As we mentioned earlier, LSA measures can be used for within-comparisons. By comparing adjacent sentences, it is possible to assess the coherence of a text: coherent texts show a high average score of semantic similarities because their sentences are generally well connected to the previous and following ones (Foltz, 2007; Foltz et al., 1998). LSA can also be used for between-comparisons. The learner essay, or any of its sentences, can be compared to a course text (Lemaire and Dessus, 2001) to a reference text (Kintsch et al., 2007) or to an ideal answer (Graesser et al., 2007). High similarities are indications of a good coverage of the reference text by the learner.

4 Automated text assessment systems: an example of genealogy

As an example, we introduce some of the systems we devised and implemented during the last ten years. First, these systems are *environments* enabling learners to freely

proceed with their tasks and request quick feedback on demand. Second, they intend to foster self-regulated learning, they are based on *several loops* learners can freely enter or leave. This section introduces the different free-text assessment systems we designed and implemented and Table 1 depicts the main functionalities of all the tools.

4.1 RAFALES, a reading-based system

RAFALES (*Recueil Automatique Favorisant l'Acquisition d'une Langue Etrangère et/ou de Spécialité*, or Automated Readings for L2 and specialist language) is an LSA-based prototype working as a digital schoolbook that autonomously chooses a sequence of texts in order to optimise the learning rate of a learner (university student) in a domain-specific foreign language (Zampa and Lemaire, 2002). The idea that learning a second language is essentially based on the exposure to the language, and not only to the explanation of the rules of that language, is recognised by researchers in second language acquisition (Krashen, 1981). The assumption behind the design of *RAFALES* is that, usually, students use schoolbooks to learn a foreign language or domain-specific foreign language which knowledge level is not tailored to them because there are few books about such a specific content.

Table 1 A genealogy of some free-text assessment systems developed in our laboratories.

Loop/system	*Reading loop*	*Writing loop*	*Interaction/feedback loop*
RAFALES	Provide texts to read neither too far nor too close to the learner's model.	N/A	N/A
Apex 1.0	N/A	Write a summary of the course from a text read beforehand (without any assessment or retrieval mechanism).	Topic coverage, outline, and inter-sentence coherence.
Resum'Web	N/A	Write a summary of the course from a text read beforehand (without any assessment or retrieval mechanism)	Prompts which summarisation strategies are used (machine's vs. learner's viewpoint).
Apex 2.0	Read a course retrieved from keywords (search engine query).	Write a summary of the texts.	Prompts which texts are understood or not (machine's vs. learner's viewpoint).
Pensum	Read a set of documents.	Write a synthesis from the read documents.	Prompts inter-sentence coherence, off-topic (synthesis), and course topic coverage. The learner has a control over the feedback.

Learners have a single task to perform with *RAFALES*: they have to read texts, which are chosen from a large knowledge database according to a student model. As a typical intelligent tutor, *RAFALES* has three major components: a pedagogical module, a domain

expert module and a student model (Wenger, 1987), all three modules relying on LSA. The student model and the knowledge database are 300-dimensional spaces derived from texts. The pedagogical module selects and sorts the texts to the learner out, according to the semantic distances between the student model and the knowledge database, thus creating a personalised and adaptive book.

Exposed to sequences of words in a random fashion, a learner would certainly build knowledge in the same way a child learns new words by reading various books. However, the process of learning could be speeded up by selecting the right sequence of words given the current state of learner entities. Therefore, the problem is to know which text has the highest chance of enlarging the part of the semantic space by learner entities. We call optimal proximity for acquisition (OPA) the semantic distance (between the learner's knowledge and the stimuli to provide) that optimises acquisition. This OPA is a model of the zone of proximal development (Vygotsky, 1962), selected texts are neither too far from nor too close to the student model. Once the learner has read the text, it is added to the student model and the semantic space is calculated. *RAFALES* selects a new text, and so on. The process (and so the course text generated) ends when the student model matches the expert model.

It is worth noting that *RAFALES* is a system that only allows reading activities, thus there is room to enable the learner to write free texts to be automatically assessed.

4.2 Apex 1.0 and Resum'Web, two writing-centred systems

Lesson and lecture content often 'evaporates' as soon as the learners have left the classroom or the distance learning course (Hübner et al., 2006), because they lack of opportunities to monitor their understanding of the course, and to detect possible misunderstandings. Essay writing may be the occasion to give learners complex activities that let them build knowledge instead of engaging in rote learning.

In such systems, on the one hand learners are committed in a writing activity, and on the other hand the semantic content of their production is automatically assessed. *Apex 1.0* (Lemaire and Dessus, 2001) is such a system also built on top of LSA, which provides feedback to a learner essay with respect to the text of a course (tagged by the teacher beforehand). Three kinds of assessments are available on demand: content-related, about the outline and the coherence of a learner essay.

Apex 1.0 assesses a learner essay with respect to a reference text, which basically corresponds to a course material. This raw text has to be marked up by the teacher so that it is divided into topics, and each topic is divided into notions. First level titles should begin with #T (for topic) whereas second level titles should begin with #N (for notion). Only second level titles are followed by a paragraph. A notion can belong to several topics. However, the text of such a notion is written only once. In case of cross-references, the notion title should begin with #S. For instance, 'Introduction to the solar system' is defined in the topic 'The solar system' and referred to in the topic 'Eclipses'.

Learners are provided with a list of topics. They select the one they wish to work on and write their knowledge about that topic in a text editor. At any time, they can get an evaluation of their essays and rewrite their text accordingly. That iterative process of writing and getting assessment is the core of the way *Apex 1.0* helps learning. It uses an LSA semantic space to generate assessments on a learner text, given a topic. Several LSA-based assessments are provided:

- *Content-based assessment: Apex 1.0* computes the semantic similarity between the essay and each of the paragraphs of the topic. A low value indicates that both texts are not semantically related; the learner therefore did not cover well that topic. Conversely, a high value would indicate a good coverage of the topic. Appropriate messages are given to the learner for each paragraph. Examples are: 'The notion 'Solar eclipses' was well covered' or 'The notion 'Lunar eclipses' was poorly covered'. An online version of *Apex 1.0* was also designed in order to give a real-time assessment, while the learner is typing in the essay. The learner can visualise the degree of coverage of each notion on a scale.
- *Outline-based assessment: Apex 1.0* attempts to predict the structure of the essay by identifying which notion best corresponds to each paragraph of the essay. In case a similarity is too low, the system warns the learner that no notion is close enough to that paragraph. Using that tool, learners are aware of the way the system 'understands' their essay. They can also have an idea of the overall structure of their essay: no notions should be missing but a notion should not be discussed in many places either.
- *Coherence-based assessment: Apex 1.0* computes the semantic similarity between all pairs of adjacent sentences within a paragraph. A low value indicates that the current topic has suddenly changed, which is probably not what the learner intended. A message at this level looks like 'There is a big conceptual break between the two following sentences of your essay: ...'.

Resum'Web (Lemaire et al., 2005; Mandin, 2010) is an LSA-based system aiming at improving the quality of learners' summaries through a better use of strategies involved in this activity. *Resum'Web* focuses on two kinds of strategies:

1 the identification of important sentences

2 the application of macrorules (Kintsch and van Dijk, 1978; van Dijk and Kintsch, 1983).

Thus, two computational and cognitive models are implemented to deliver a feedback on the accuracy of learners' answers about two tasks (selection of the five most important sentences of the source text and the identification of strategies used to write each summary sentence). The learners are then prompted to assess their own strategies and to check whether or not they match with those diagnosed by *Resum'Web*.

These two systems are a step further in the SRL development of learners: they first let learners write texts (thus they have a better view on what they learnt compared to read-only systems) and also let them reflect on their own writing strategies. However, they do not involve learners in two activity loops, with an assessment on each, as the next systems do.

4.3 Apex 2.0 and Pensum, two reading/writing-based systems

As the previous systems, *Apex 2.0* and *Pensum* are LSA-based systems, but the activity they propose is divided up in reading and writing loops, the latter being automatically assessed.

Apex 2.0 (Dessus and Lemaire, 2002) engages the learner in two different loops. The first one lets him or her choose a subject content to study, from a query composed of one

or more keyword(s). Then the system delivers a set of documents semantically close to this query. At any time the learner can write out a summary of the document he or she has read and understood and can get an automated feedback on this summary. In so doing, the learner can revise their exams and enter a self-regulated workflow: reading-writing-feedback-revision, which could lead to a better understanding of the chosen subject content.

The aim of *Pensum* (Dessus et al., 2010; Villiot-Leclercq et al., 2010) is to help learners in their course learning. It is intended to be used on an LLL platform, potentially in conjunction with other educational applications, each application (including *Pensum*) being available as a widget. In *Pensum*, the learners can read their courses (reading loop) and decide when they want to synthesise them (writing loop). They can ask for automatic feedback on their production at any time. Then, the system delivers content assessment about three points:

1 the coherence between consecutive sentences

2 whether a sentence of the synthesis is off-topic (i.e., not related to any sentence of the course)

3 the course sentences covered in the synthesis (i.e., likely to be integrated).

The response to each of those types of feedback is rather straightforward in terms of user response.

The aforementioned two types of external feedback would then fall into some of the "principles of good feedback practice" (Nicol and Macfarlane-Dick, 2006), as it "encourages teacher and peer dialogue around learning" and allows the learner to "make some kind of response to complete the feedback loop" (Sadler, 1998).

5 Lessons learned

Some lessons have been learned from the validation of these systems in laboratory or more ecological settings (i.e., distance learning courses, exams). All these systems have been tested with tenth grade pupils (*Resum'Web*) and university students (the other ones). We cannot report in this paper the quantitative results of each validation study, we shall nevertheless list the most important challenges with regard to free-text assessment.

- *Which status for feedback?* Our view concerning feedback is in accordance with Sadler's view of formative assessment according to which learners should be able to "make the connections between the feedback and the work they produce, and how they can improve their work in the future" [Sadler, (1998), p.78]; or so it would, if the feedback were 100% reliable. But, as stated above, automatic analysis of the semantic level of free text can hardly claim such reliability. We therefore decided to extend the concept of 'didactic triangulation' proposed in Blanchard et al. (2009) by working on the type of feedback proposed in accordance with the limits of the NLP tools used (namely LSA in our case), rather than focusing on almost perfectly reliable language technologies. In order to improve the adoptability of the system by both tutors and learners by improving the readability of the aims and capacities of the system (Murray and Barnes, 1998) and thus allow them to make better use of the system (Bax, 2003), *Pensum* allows the user to question the feedback provided (both

references come from the field of computer assisted language learning, but the notions evoked in these papers are adaptable to a broader range of educational systems). The learner can even justify it by linking sentences from the course and from the synthesis. The integration of such functionalities in the core of *Pensum* clarifies the status of the feedback, which is not to be considered definitive, but rather hints on possible problems concerning the synthesis and an aid to the learner in terms of indicating their coverage of the concept of the course learnt. An extensive history of the successive states of the synthesis is meant to avoid extreme tediousness in the use of this functionality (as a feedback rejected by the user will be displayed as questioned for subsequent versions of the synthesis, until the learner changes his or her judgement) and allow the user to build on their former responses to the feedback. This improves the quality of the feedback itself if we consider the suggestion of Nicol and Macfarlane-Dick (2006, p.208): "Good quality external feedback is information that helps students troubleshoot their own performance and self-correct: that is, it helps students take action to reduce the discrepancy between their intentions and the resulting effects". Indeed, the feedback items can be considered as questions asked to the user. The answers to these questions are to help learners to ascertain the validity of their synthesis, modify it or keep track of the elements of the course they have covered even when they are not mentioned in their synthesis (rejection of coverage feedback can be either stored as an explicit link to the synthesis or as an annotation of the course in terms of irrelevant sentences).

- *How to improve the guidance functionalities?* Pointing out problems related to the written productions is not enough for learners. The learners also expect guidance on how to fix them. Learners were given satisfaction questionnaires after using *Resum'Web* or *Pensum*, indicating they would like more advanced functions prompting them what they should do (e.g., an example showing how to increase the coherence between sentences). Thus, these systems could be improved in providing automatically generated examples (Capus and Tourigny, 1998, 2003) or in fostering knowledge-creation discourses (Scardamalia and Bereiter, 2006; van Aalst, 2009).

- *Which feedback is effective?* Kulhavy and Stock (1989), cited by Shute (2008) showed that effective feedback provides the learner with two types of informational feedback: verification (indicating if the answer is correct or not) and elaboration (helping learners find the correct answer with a set of relevant cues). They argued that an efficient feedback should mix verification and elaboration feedback (address the topic, discuss the particular error, give gentle guidance). The effectiveness of an 'elaborated feedback' is confirmed by the review of studies led by Shute (2008), showing that efficient feedback aims at helping learners to improve their answers rather than indicating if the answer is correct or not and being specific on the task, especially for supporting learners in retention or comprehension tasks. The systems presented above (*RAFALES* excepted) all propose a kind of 'elaborated feedback' more centred on the way learners can act upon their writing rather than take the feedback as granted.

- *Teachers cannot be replaced*. Delivering just-in-time feedback is useful for learners, even though not always fully reliable. For instance, the comparison of the computational cognitive models implemented in our software against human judgements (e.g., Mandin et al., 2007) showed weak to pretty strong correlations;

and sometimes even wrong feedback. However, an experiment involving pupils with *Resum'Web* showed an increase of the identified strategies as long as *Resum'Web* was used, while the quality of work, assessed by teachers, was stable (Mandin, 2010). Of course, the computer cannot be the sole evaluator of the written production if the goal is to provide as reliable feedback as possible: the teacher is needed, as confirmed by one of our studies. After using a first version of *Pensum*, which allowed no interaction with a teacher through the system, nor the possibility for the learner to question the feedback, a large number of students requested to have interactions with a teacher or tutor, either for their lack of confidence towards the system or to have more information on the feedback.

- *How to improve the models?* Without going into algorithmic details of the models, we suggest some design modifications to improve such automated text assessment systems. First, recall that the semantic analysis heavily depends on the proper spelling of words. However, students make a lot of spelling errors, causing a decrease in performance. It is therefore necessary to use a spell and grammatical checker beforehand. Second, we can test a large range of models in which a large range text features may vary (e.g., type, length) (Lemaire et al., 2005); and whose results may be better than those provided by LSA (Fernandez et al., 2008).

6 Conclusions

The systems we presented in this paper all have in common two main features. First, they are aimed at fostering learners' comprehension on course texts. Second, they share the same methodological background about free-text assessment (i.e., LSA-based), but have different roles with regard to the activity loops and the SRL processes the learners are engaged in.

RAFALES proposes read-only activities and no specific SRL activities. *Apex 1.0* and *Resum'Web*, in involving learners in writing activities as well as comparing their own with that of the machine, fosters a first level of self-regulation; nonetheless, the reading activities are poorly managed in these two systems. A step further, *Apex 2.0* and *Pensum* add to the previous ones a richer reading loop, and, most importantly, *Pensum* even proposes a way for the learner to act upon the feedback of the machine.

Like Mory (2004), we assume in this paper that the development of free-text assessment techniques opens a lot of possibilities to make feedback more adaptive, more effective, and more centred on higher-level activities (e.g., SRL, summarisation, strategy identification).

Acknowledgements

The design and development of *Resum'Web has* been supported by a grant from the French Research Ministry under an 'Ecole et sciences cognitives' research project. Pensum is designed and developed as part of Language Technologies and Lifelong Learning (LTfLL), a FP 7 EC-funded research project.

References

Antoniadis, G. (2004) 'Les logiciels d'apprentissage des langues peuvent-ils ignorer le TAL?', *Les cahiers de l'APLIUT*, Vol. 23, No. 2, pp.82–97.

Bangert-Drowns, R.L., Hurley, M.M. and Wilkinson, B. (2004) 'The effects of school-based writing-to-learn interventions on academic achievement: a meta-analysis', *Review of Educational Research*, Vol. 74, No. 1, pp.29–58.

Bax, S. (2003) 'CALL – past, present and future', *System*, Vol. 31, No. 1, pp.13–28.

Blanchard, A., Kraif, O. and Ponton, C. (2009) 'Mastering noise and silence in learner answers processing: simple techniques for analysis and diagnosis', *CALICO Journal*.

Burstein, J., Kukich, K., Wolff, S., Lu, C. and Chodorow, M. (1998) 'Computer analysis of essays', *NCME Symposium on Automated Scoring*, Montréal.

Capus, L. and Tourigny, N. (1998) 'Learning summarization by using similarities', *Computer Assisted Language Learning*, Vol. 11, No. 5, pp.475–488.

Capus, L. and Tourigny, N. (2003) 'A case-based reasoning approach to support story summarization', *International Journal of Intelligent Systems*, Vol. 18, pp.877–891.

Dessus, P. (2009) 'An overview of LSA-based systems for supporting learning and teaching', in Dimitrova, V., Mizoguchi, R., du Boulay, B. and Graesser, A. (Eds.): *Artificial Intelligence in Education: Building Learning Systems that Care: From Knowledge Representation to Affective Modelling (AIED2009)*, pp.157–164, IOS Press, Amsterdam.

Dessus, P. and Lemaire, B. (2002) 'Using production to assess learning: an ILE that fosters self-regulated learning', in Cerri, S.A., Gouardères, G. and Paraguaçu, F. (Eds.): *Intelligent Tutoring Systems (ITS 2002)*, pp.772–781, Springer, Berlin.

Dessus, P., Trausan-Matu, S., Mandin, S., Rebedea, T., Zampa, V., Dascalu, M. and Villiot-Leclercq, E. (2010) 'Assessing writing and collaboration in learning: methodological issues', Paper presented to the *Workshop 'Analysing the Quality of Collaboration in Task-oriented Computer-mediated Interactions', held in conjunction to the 9th International Conference on the Design of Cooperative Systems (COOP 2010)*, Aix-en-Provence.

Ericsson, P. and Haswell, R. (2006) *Machine Scoring of Student Essays: Truth and Consequences*, Utah State University Press, Logan.

Fernandez, S., Mandin, S., Torres, J-M. and Velazquez, P. (2008) 'Les systèmes de résumé automatique sont-ils vraiment de mauvais élèves?', in Heiden, S. and Pincemin, B. (Eds.): *9e Journées Internationales d'Analyse Statistique des Données Textuelles (JADT 2008)*, pp.469–481, ENS Lettres & Sciences Humaines, Lyon.

Fischer, G. (2001) 'Lifelong learning and its support with new media', in Baltes, P.B. and Smelser, N.J. (Eds.): *International Encyclopedia of the Social & Behavioral Sciences*, pp.8836–8840, Elsevier, Oxford.

Foltz, P.W. (2007) 'Discourse coherence and LSA', in Landauer, T.K., McNamara, D., Dennis, S. and Kintsch, W. (Eds.): *Handbook of Latent Semantic Analysis*, pp.167–184, Erlbaum, Mahwah.

Foltz, P.W., Kintsch, W. and Landauer, T.K. (1998) 'The measurement of textual coherence with latent semantic analysis', *Discourse Processes*, Vol. 25, Nos. 2–3, pp.285–307.

Foltz, P.W., Laham, D. and Landauer, T.K. (1999) 'Automated essay scoring: applications to educational technology', *Proc. Int. Conf. ED-MEDIA '99*, Seattle.

Graesser, A.C., Penumatsa, P., Ventura, M., Cai, Z. and Hu, X. (2007) 'Using LSA in AutoTutor: learning through mixed-initiative dialogue in natural language', in Landauer, T.K., McNamara, D., Dennis, S. and Kintsch, W. (Eds.): *Handbook of Latent Semantic Analysis*, pp.243–262, Erlbaum, Mahwah.

Graham, S. and Perin, D. (2007) 'A meta-analysis of writing instruction for adolescent students', *Journal of Educational Psychology*, Vol. 99, No. 3, pp.445–476.

Hübner, S., Nückles, M. and Renkl, A. (2006) 'Fostering the cycle of self-regulation in writing learning protocols', in Clarebout, G. and Elen, J. (Eds.): *Avoiding Simplicity, Confronting Complexity: Advances in Studying and Designing (Computer-based) Powerful Learning Environments*, pp.155–164, Sense Publishers, Rotterdam.

Kakkonen, T., Myller, N. and Sutinen, E. (2004) 'Semi-automatic evaluation features in computer-assisted essay assessment', *7th IASTED International Conference on Computers and Advanced Technology in Education (CATE 2004)*, Kauai, pp.456–461.

Kakkonen, T., Myller, N. and Sutinen, E. (2006) 'Applying part-of-speech enhanced LSA to automatic essay grading', *4th IEEE International Conference on Information Technology: Research and Education (ITRE 2006)*, Tel Aviv, Israel.

Kintsch, E., Caccamise, D., Franzke, M., Johnson, N. and Dooley, S. (2007) 'Summary street: computer-guided summary writing', in Landauer, T.K., McNamara, D., Dennis, S. and Kintsch, W. (Eds.): *Handbook of Latent Semantic Analysis*, pp.263–277, Erlbaum, Mahwah.

Kintsch, W. and van Dijk, T.A. (1978) 'Toward a model of text comprehension and production', *Psychological Review*, Vol. 85, No. 5, pp.363–394.

Klein, P.D. (1999) 'Reopening inquiry into cognitive processes in writing-to-learn', *Educational Psychology Review*, Vol. 11, No. 3, pp.203–270.

Koper, R. (2004) 'Use of the semantic web to solve some basic problems in education', *Journal of Interactive Media in Education*, Vol. 6, pp.1–23.

Krashen, S.D. (1981) *Second Language Acquisition and Second Language Learning*, Pergamon Press, Oxford.

Kulhavy, R.W. and Stock, W. (1989) 'Feedback in written instruction: the place of response certitude', *Educational Psychology Review*, Vol. 1, No. 4, pp.279–308.

Landauer, T.K., McNamara, D.S., Dennis, S. and Kintsch, W. (2007) *Handbook of Latent Semantic Analysis*, Erlbaum, Mahwah.

Langer, J.A. and Applebee, A.N. (1987) *How Writing Shapes Thinking: A Study of Teaching and Learning*, National Council of Teachers of English, Urbana, IL.

Lemaire, B. and Dessus, P. (2001) 'A system to assess the semantic content of student essays', *Journal of Educational Computing Research*, Vol. 24, No. 3, pp.305–320.

Lemaire, B., Mandin, S., Dessus, P. and Denhière, G. (2005) 'Computational cognitive models of summarization assessment skills', in Bara, B.G., Barsalou, L. and Bucciarelli, M. (Eds.): *Proceedings of the 27th Annual Conference of the Cognitive Science Society (CogSci' 2005)*, Erlbaum, Mahwah, pp.1266–1271.

Lin, X. (2001) 'Designing metacognitive activities', *Educational Technology Research & Development*, Vol. 49, No. 2, pp.23–40.

Linn, M.C. (1996) 'Cognition and distance learning', *Journal of the American Society for Information Science*, Vol. 47, No. 11, pp.826–842.

Mandin, S. (2010) 'Computational cognitive models of summarizing: TEL environment experiment on high school students', *Junior Researchers of EARLI Conference (JURE 2010)*, Frankfurt.

Mandin, S., Lemaire, B. and Dessus, P. (2007) 'Modeling summarization assessment strategies with LSA', in Wild, F., Kalz, M., van Bruggen, J. and Koper, R. (Eds.): *Proc. First European Workshop on Latent Semantic Analysis in Technology Enhanced Learning*, Open University of the Netherlands, Heerlen, pp.20–21.

Mory, E.H. (2004) 'Feedback research revisited', in Jonassen, D.H. (Ed.): *Handbook of Research on Educational Communications and Technology*, 2nd ed., pp.745–783, Erlbaum, Mahwah.

Murray, L. and Barnes, A. (1998) 'Beyond the 'wow' factor – evaluating multimedia language learning software from a pedagogical viewpoint', *System*, Vol. 26, No. 2, pp.249–259.

Nicol, D.J. and Macfarlane-Dick, D. (2006) 'Formative assessment and self-regulated learning: a model and seven principles of good feedback practice', *Studies in Higher Education*, Vol. 31, No. 2, pp.199–218.

Page, E. (1966) 'The imminence of grading essays by computer', *Phi Delta Kappan*, Vol. 47, pp.238–243.

Sadler, D.R. (1998) 'Formative assessment: revisiting the territory', *Assessment in Education: Principles, Policy & Practice*, Vol. 5, No. 1, pp.77–84.

Scardamalia, M. and Bereiter, C. (2006) 'Knowledge building: theory, pedagogy, and technology', in Sawyer, R.K. (Ed.): *The Cambridge Handbook of the Learning Sciences*, pp.97–115, Cambridge University Press, Cambridge.

Shute, V.J. (2008) 'Focus on formative feedback', *Review of Educational Research*, Vol. 78, No. 1, pp.153–189.

van Aalst, J. (2009) 'Distinguishing knowledge-sharing, knowledge-construction, and knowledge-creation discourses', *Computer-Supported Collaborative Learning*, Vol. 4, pp.259–287.

van Dijk, T.A. and Kintsch, W. (1983) *Strategies of Discourse Comprehension*, Academic Press, London.

Villiot-Leclercq, E., Mandin, S., Dessus, P. and Zampa, V. (2010) 'Helping students understand courses through written syntheses: an LSA-based online advisor', *Tenth IEEE Int. Conf. on Advanced Learning Technologies (ICALT 2010)*, Sousse, Tunisia.

Vovides, Y., Sanchez-Alonso, S., Mitropoulou, V. and Nickmans, G. (2007) 'The use of e-learning course management systems to support learning strategies and to improve self-regulated learning', *Educational Research Review*, Vol. 2, No. 1, pp.64–74.

Vygotsky, L.S. (1962) *Thought and Language*, MIT Press, Cambridge.

Wade-Stein, D. and Kintsch, E. (2004) 'Summary street: interactive computer support for writing', *Cognition and Instruction*, Vol. 22, No. 3, pp.333–362.

Wenger, E. (1987) *Artificial Intelligence and Tutoring Systems*, Morgan Kaufmann, Los Altos.

Williamson, D.M., Mislevy, R.J. and Bejar, I.I. (Eds.) (2006) *Automated Scoring of Complex Tasks in Computer-based Testing*, Erlbaum, Mahwah.

Wresch, W. (1993) 'The imminence of grading essays by computer – 25 years later', *Computers and Composition*, Vol. 10, No. 2, pp.45–58.

Zampa, V. and Lemaire, B. (2002) 'Latent semantic analysis for user modeling', *Journal of Intelligent Information Systems*, Vol. 18, No. 1, pp.15–30.