Assistance informatique à la correction de copies

Philippe Dessus, Benoit Lemaire

Paru en 2004 In E. Gentaz & P. Dessus (Eds.). *Comprendre les apprentissages, Sciences cognitives et éducation* (pp. 205-220). Paris : Dunod.

Assistance informatique à la correction de copies

Philippe Dessus*± & Benoît Lemaire*

* Laboratoire des sciences de l'éducation
Université Pierre-Mendès-France
BP 47
38040 Grenoble CEDEX 9

± IUFM de Grenoble
30 av. Marcellin-Berthelot
38100 Grenoble

Résumé

Corriger des copies est, pour l'enseignant, une des tâches les plus coûteuses en temps et en charge cognitive, aussi, de nombreuses méthodes, qu'elles soient manuelles (questionnaires à choix multiple) ou informatiques, ont été élaborées pour l'assister. Nous passerons en revue ces dernières, après avoir montré que les méthodes utilisées sont principalement calquées sur celles mises en œuvre par un correcteur humain. Nous distinguons deux types d'aides : une aide qui se centre sur les aspects de surface du texte (orthographe, grammaire, longueur des mots, etc.), et une aide qui se centre sur des aspects liés au contenu traité dans la copie. La première partie détaille les processus psychologiques en œuvre dans cette activité, la deuxième partie passe en revue les principaux systèmes informatiques, tout d'abord ceux qui réalisent un traitement de surface, ensuite ceux qui se centrent sur le contenu traité.

INTRODUCTION

L'objet de ce chapitre est de passer en revue les différents moyens dont l'enseignant peut disposer pour être assisté, par des aides informatiques, dans une tâche difficile : celle de la correction de copies. L'enseignant a toujours essayé de diminuer sa charge de travail concernant ce domaine, en utilisant divers moyens ou méthodes non informatiques (barème, questionnaires à choix multiple, questions fermées, etc.). La popularisation de l'informatique aux différents niveaux de l'enseignement ouvre de nouvelles perspectives, que nous allons détailler ici. Il est intéressant de noter que, si les effets des aides informatiques sur l'activité de production écrite ont été largement étudiés [1], ceux concernant l'évaluation de cette même activité par l'informatique ont été délaissés. Même si les moyens informatiques décrits ci-dessous ne sont pas encore aisément utilisables au quotidien par les enseignants, il en existe, sur Internet notamment, des versions de démonstration qui permettent de se faire une idée de leur intérêt.

En 1999, un programme informatique, *e-rater*, a systématiquement remplacé un des deux correcteurs humains des dissertations de candidats au test GMAT (*Graduate Management Admission Test*), un test d'admission aux grandes écoles de commerce internationales (MBA). En cas de désaccord de plus d'un point entre les deux notes, une troisième correction, humaine, est réalisée. Il est aisé de comprendre combien la charge de correction (500 000 copies par an) est allégée, si l'on peut ainsi remplacer une partie de ce travail par une machine [2]. Précisons maintenant ce qu'on entend par les différents termes employés dans le titre de ce chapitre.

QUELQUES DEFINITIONS

Par *correction*, nous entendons production de l'enseignant visant à accompagner une décision concernant l'élève ayant produit cette copie. La décision peut concerner des événements à court terme (*e.g.*, évaluer les compétences de l'élève pendant ou à la suite d'un apprentissage), ou à plus long terme (*e.g.*, prendre une décision d'orientation). Les compétences de l'élève peuvent être soit des compétences d'écriture (évaluation *de* l'écrit), ou encore des compétences dans un domaine précis (évaluation *par* l'écrit). Par *copies*, nous entendons productions écrites en réponse à une consigne (*e.g.*, décrire, résumer, argumenter, etc.), ce qui nous fera écarter de cette revue la réponse à des questionnaires à choix multiple. De plus, nous ne nous centrons pas ici sur un contenu spécifique, la seule condition étant que

les contenus traités aient une forme textuelle, ce qui les rend facilement analysables par des moyens informatiques.

À propos d'*assistance*, il nous faut notamment préciser qu'il n'est pas question de proposer des outils informatiques qui donneraient une correction à l'étudiant en dehors de toute interaction avec un enseignant (qu'elle soit en présence ou à distance). French [3] signale trois principales objections à cette idée : *une objection humaniste*, qui pose que seul un humain peut opérer certains choix liés à ce travail ; une *objection défensive*, qui signale qu'il est nécessaire qu'un être humain règle les « cas limites » (étudiants qui ne respectent pas le contrat didactique, hors-sujet, etc.) ; une *objection de contenu*, qui questionne le fait que les indicateurs évalués par l'ordinateur sont bien les indicateurs importants. Si certains paramètres ne pourront être traités par la machine avant longtemps (*e.g.*, le style), la plupart des pratiques de correction de copies ont déjà fait l'objet d'une informatisation.

Ce chapitre comprend deux parties principales. La première traite des processus cognitifs engagés dans l'activité de correction de copies. La deuxième recense les principales méthodes d'assistance à la correction de copies. Il nous semble en effet que de tels moyens informatiques à l'assistance à la correction de copies ne sont pas que des curiosités académiques, des prouesses techniques d'informaticiens, mais bien des travaux qui nous permettent de mieux comprendre les processus cognitifs humains de ce domaine.

PROCESSUS COGNITIFS DANS L'ACTIVITE DE CORRECTION DE COPIES

Avant de décrire les différentes aides informatiques à la correction de copies, il est utile, d'une part de se demander quels sont les processus cognitifs engagés par un enseignant dans une telle activité et, d'autre part, de voir de quelle manière ils peuvent être simulés et/ou assistés. Ainsi, il est possible de proposer des aides à l'évaluation de copies qui soient proches des processus évaluatifs humains. Notons aussi que, paradoxalement, le fait d'élaborer des systèmes d'assistance à une activité permet de mieux comprendre de quelle manière l'on réalise cette activité. Par exemple, l'enseignant corrigeant une série de copies est susceptible d'effectuer « à la main » un certain nombre d'opérations également réalisées dans les systèmes informatiques décrits plus bas (*e.g.*, un classement des copies, par rapport à une norme). Ce classement s'apparente à ce que l'*Intelligent Essay Assessor* (*voir plus bas*) réalise dans sa méthode « *gold standard* ».

Que fait l'enseignant quand il corrige des copies ?

L'activité d'évaluation de copies a fait l'objet, dans le champ de recherche sur l'évaluation, d'assez rares travaux. Les raisons pour lesquelles cette activité est difficile à analyser sont

nombreuses : tout d'abord, l'activité d'évaluation nécessite que l'enseignant évalue à la fois : — la nature possible des activités de l'élève ayant produit la copie (a-t-il bien compris le contenu ? a-t-il bien résumé ou rendu compte de ce contenu ?) ; — la qualité de la copie, à la fois en lien avec ce qu'il en attendait et la qualité moyenne des copies des élèves ; — la qualité de ses propres performances en tant qu'évaluateur (ne suis-je pas trop, ou pas assez sévère ?). De plus, l'enseignant ne se contente en général pas de lire, mais également de proposer des alternatives ou des corrections au texte qu'il propose, ce qui nécessite une charge cognitive supplémentaire. En d'autres termes, lire une copie pour l'évaluer nécessite bien d'autres activités que de la lire pour la comprendre [4].

Ensuite, il est délicat, à cause de ces différentes activités complexes, de demander *en plus* à l'enseignant qu'il donne, en direct, pendant sa correction de copies, des renseignements sur la manière dont il s'y prend. Il nous semble que l'activité de révision de textes (*i.e.*, amélioration du texte par son auteur, voir par exemple [5]) partage de nombreux points avec celle de l'évaluation (même si, dans cette dernière, s'ajoute la production d'une note). En effet, le lecteur d'un texte à des fins de révision ou d'évaluation le lit pour le comprendre, pour détecter d'éventuels problèmes, pour proposer d'éventuelles alternatives. Un évaluateur de copies aurait, lui, un but principal, comprendre la copie, et d'autres buts de haut niveau, comme critiquer le texte du point de vue de son efficacité, de son style, ou bien encore la production de suggestions améliorant ces derniers paramètres.

C'est à notre connaissance la revue de Schriver [4] qui nous permet de mieux préciser ces liens entre l'activité de révision de productions écrites et celle de correction. Une perspective de communication permet d'avancer que deux types d'informations sont récupérés lors de l'évaluation d'une production écrite : une information à propos de la qualité globale du texte produit, et une information sur la manière dont l'audience pourra réagir au texte. Ainsi, le classement des méthodes d'évaluation se réalise en trois parties :

— *Les évaluations centrées sur le texte*, dans lesquelles l'évaluateur (qu'il soit un enseignant ou un ordinateur) examine le texte au regard d'un ensemble de caractéristiques, et évalue la qualité de ce dernier via un ensemble de principes prescriptifs (*e.g.*, il ne peut y avoir des paragraphes plus longs qu'une demi-page). Ici, la réponse directe d'un éventuel lecteur n'est pas envisagée.

— *Les évaluations centrées sur le jugement d'expert*, dans lesquelles l'évaluateur, possédant en principe un haut niveau de connaissances du contenu traité dans la production écrite, juge cette dernière. Le jugement va donc se centrer nécessairement sur des aspects plus liés au contenu.

— *Les évaluations centrées sur le lecteur*, font part au producteur de l'écrit du jugement du public prévu pour ce dernier, par le biais d'annotations. Dans ce cas (*e.g.*, écriture d'un journal scolaire), l'enseignant joue le rôle d'un lecteur extérieur à la classe.

Dans la suite de cette revue, nous laisserons de côté le troisième type d'évaluation. En effet, dans un contexte scolaire, il est rarement fait mention de l'audience supposée de l'écrit : dans la majorité des écrits scolaires, l'audience prévue est l'enseignant lui-même. Passons en revue quelques méthodes des deux premières catégories.

Évaluation centrée sur le texte

L'enseignant, lorsqu'il corrige des copies d'élèves ou d'étudiants, réalise un certain nombre d'évaluations centrées sur le texte. Les plus courantes sont liées à la calligraphie, à l'orthographe (erreurs d'orthographe, de ponctuation, d'accents, etc.), à la grammaire (conjugaison, différents types d'accord), à la forme de la présentation (longueur des paragraphes), et à la cohérence interparagraphes (problèmes de transition, etc.). Ces différentes évaluations, comme on l'a montré [5], sont réalisées en parallèle, en tant que processus distincts, et c'est pour cela que corriger (ou réviser) un texte est une activité cognitive importante. Le processus de ces évaluations est maintenant bien documenté, aussi nous ne nous y attarderons pas, mais noterons qu'il est facilité depuis l'apparition des logiciels de traitement de textes.

Évaluation centrée sur le jugement d'expert

D'autres méthodes se sont développées en réaction au développement de la méthode précédente, plus analytique. Elles conviennent pour évaluer la qualité littéraire d'un texte (évaluation *de* l'écrit), mais également la qualité de son contenu (évaluation *par* l'écrit)**.** L'idée sous-jacente à ces méthodes est que comprendre un texte, c'est être capable de le résumer. Même si l'on a pu montrer que le lien entre compréhension et résumé n'était pas systématique, il reste que la capacité d'un élève (au moins de niveau du collège) à résumer un texte est un bon prédicteur de ses connaissances sur le domaine du texte. D'ailleurs, une grande partie des examens de fin d'année sont bien, dans une certaine mesure, des sollicitations à produire des résumés de certaines parties de cours. Trois méthodes peuvent être citées.

Le classement par impression générale. L'évaluateur range chaque copie de la meilleure à la moins bonne, selon l'impression générale qu'il a eue à sa lecture [4].

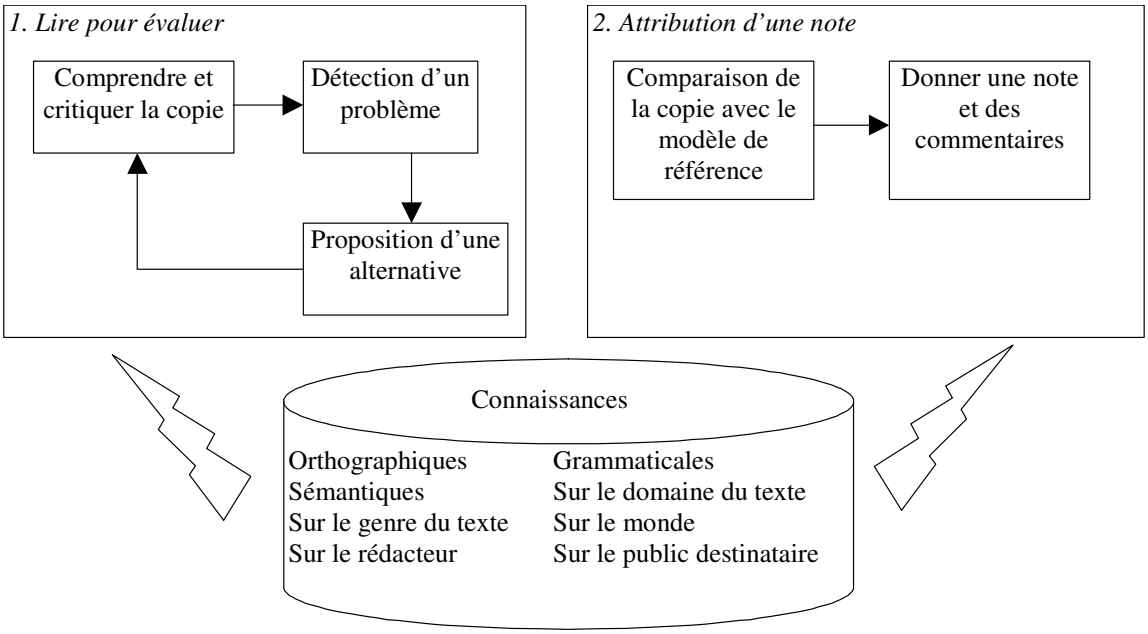
Le score holistique. Avant l'évaluation, le correcteur a été entraîné à cette activité (pendant environ 2 h) par la lecture d'un certain nombre de copies déjà évaluées et classées en plusieurs catégories (par exemple, insuffisant, moyen, bon, très bon). Ensuite, évaluer chaque copie revient à la ranger dans l'une des catégories. Ce type d'évaluation [6], comme son nom l'indique, amène à considérer chaque copie comme un tout, mais n'offre que peu d'intérêt si l'on considère qu'évaluer, c'est donner à l'élève des indications sur sa performance, et surtout les moyens de l'améliorer.

Le score du trait primaire ou des traits multiples (*primary trait scoring*). Cette méthode est une évaluation critériée dans laquelle on donne à l'évaluateur un critère (primaire) ou une liste de critères (multiple) de qualité d'une copie (*e.g.*, orthographe, absence de hors-sujet, style, structure), qui servent ensuite à la noter [6]. Bien évidemment, ces critères dépendent de la tâche donnée à l'élève, et doivent donc être fixés à chaque évaluation.

La correction de copies, une vue intégrée

Les méthodes d'évaluation ci-dessus ne prennent pas en compte tous les phénomènes liés à la psychologie de la correction de copies [7, 8]. Par exemple, le fait que l'enseignant examine un ensemble de copies, plutôt qu'une seule, influe sur sa notation. Il est par exemple bien connu qu'un enseignant commence par noter les premières copies d'une pile plus sévèrement et qu'il note aussi plus sévèrement une mauvaise copie suivant une bonne. On peut rendre compte de ces effets par un modèle (*voir figure 1 ci-dessous*), qui fait que l'enseignant évaluateur compare l'image qu'il a de la copie à une norme, image qu'il en attend (qui elle-même peut varier au cours de la correction de la pile de copies).

Figure 1 — Processus de correction de copies [5, p. 76-77 ; 9, p. 115]



Ainsi, l'évaluateur entre dans une première boucle, dans laquelle il lit la copie pour la comprendre et la critiquer. Deux autres sous-tâches sont tour à tour mises en œuvre, selon les problèmes rencontrés dans la copie : détection d'un problème et proposition éventuelle d'une alternative. Cette activité requiert un grand nombre de connaissances expertes, qui sont mentionnées dans le cylindre (*voir [5] pour plus de précisions*). Ensuite, une deuxième phase est mise en œuvre, qui permet d'attribuer une note à la copie. Dans la deuxième partie de ce chapitre, nous passons en revue les moyens informatiques pour remplir la même tâche de correction, avec des méthodes souvent similaires.

ASSISTANCE INFORMATIQUE A LA CORRECTION DE COPIES

Nous allons reprendre la distinction établie précédemment entre l'évaluation centrée sur le texte et centrée sur un jugement, mais en détaillant de quelle manière des moyens informatiques peuvent les remplir. Bien évidemment, ici, le jugement ne sera pas un jugement d'expert *réel*, mais un jugement simulé.

Évaluation informatique centrée sur le texte

Une manière ancienne d'évaluer en surface une production écrite est de mesurer la fréquence de ces mots [9] : en comptant les occurrences de ces mots dans un vaste corpus, on détermine leur fréquence moyenne d'apparition et ainsi leur caractère usuel ou rare. Par suite, on peut estimer la lisibilité d'un texte : il sera d'autant plus lisible qu'il contiendra des mots usuels. Ainsi, dans les années 1940, diverses formules de lisibilité de textes ont été proposées. Elles se retrouvent encore aujourd'hui dans divers logiciels de traitement de textes (*voir encadré*).

Encadré 1 — Deux formules de lisibilité d'un texte

En 1948, Dale et Chall [10] ont proposé une formule pour estimer la lisibilité d'un texte. Cette formule dépend de deux critères : la longueur des phrases et la difficulté des mots. Pour le premier, il suffit de calculer l'indice *LP*, le nombre moyen de mots par phrases. Pour le second, Dale et Chall ont défini une liste de 3 000 mots usuels et ont défini *MD*, le pourcentage de mots du texte n'appartenant pas à cette liste. Après avoir effectué de nombreux tests, ils ont abouti à la formule ci-dessous. Dale et Chall ont étalonné cet indice avec différents textes. Au deux extrémités, on trouve un score inférieur à 4,9 au niveau CM1 *(4th grade)* et un score supérieur à 10 pour le niveau Bac+4 (*College graduate*).

$$\text{Score de lisibilité de Dale et Chall} = (0{,}1579 \times MD) + (0{,}0496 \times LP) + 3{,}6365$$

La formule de Flesch est fondée, quant à elle, sur la longueur moyenne des mots et des phrases, l'idée étant que la lisibilité est d'autant plus grande que les mots comportent peu de syllabes et les phrases peu de mots. Soit *MP* le nombre moyen de mots par phrase et *SM* le nombre moyen de syllabes par mot :

$$\text{Score de lisibilité de Flesch} = 206{,}835 - (1{,}015 \times MP) - (84{,}6 \times SM)$$

Ces formules de lisibilité se centrent sur peu d'aspects de surface (nombre ou longueur de mots). D'autres travaux ont continué dans cette voie, en proposant des formules plus complexes, tenant compte d'un plus grand nombre de paramètres. C'est le cas de Page [11] qui a élaboré un programme, *PEG* (pour *Project Essay Grade*), calculant automatiquement certains occurrences d'indices du texte à évaluer pouvant être des indicateurs de sa qualité : par exemple, la longueur moyenne du texte en mots, le nombre de virgules, le nombre de connecteurs, la longueur moyenne des mots. Page a tout d'abord montré que chacun de ces indicateurs corrélait moyennement avec la qualité des textes, évaluée par des juges humains (par exemple, 0,51 pour la longueur moyenne des mots, 0,34 pour le nombre de virgules, 0,32 pour la longueur du texte). Page a également montré qu'une combinaison linéaire de ces indicateurs corrélait de manière importante avec des jugements de qualité humains (de l'ordre de 0,80).

D'autres travaux dans cette lignée (par exemple, [12]) ont également montré des corrélations du même ordre de grandeur. Ces travaux, plus sophistiqués, font correspondre certains traits de surface du texte à des caractéristiques liées à la syntaxe, la rhétorique ou le contenu du texte. Il faut aussi noter que la plupart des analyses « stylistiques » des copies d'étudiants se réalisent à partir des aspects de surface. Le principe est le même que précédemment : corréler des critères de surface avec des notes émises par des humains, de manière à pondérer chacun dans une formule générale [13]. Le principal problème lié à cette technique est qu'elle est totalement indépendante du domaine de connaissance traité dans le texte. C'est à ce problème que s'attache la deuxième catégorie de techniques, que nous exposons maintenant.

Évaluation informatique centrée sur le jugement d'experts

Ces techniques ont un point commun : une méthode d'analyse automatique de contenu permet de représenter à la fois le domaine de connaissances (*i.e.*, le cours) et la production des étudiants, puis d'effectuer une comparaison entre ces deux types de textes. Il existe plusieurs méthodes permettant cette comparaison. Une des plus populaires est l'analyse de la sémantique latente (*Latent Semantic Analysis*, voir encadré 2). Un grand nombre de logiciels ont été conçus à partir de cette méthode. S'ils partagent le même moteur, ils se distinguent les uns des autres par le traitement réalisé sur les textes. Il est à noter que ce traitement est en grande partie similaire aux différentes méthodes d'évaluation « humaines » détaillées plus haut. D'ailleurs, les valeurs de corrélations entre les notes humaines et celles obtenues par ce type de logiciel sont assez élevées (entre 0,6 et 0,85).

Encadré 2 — Présentation de l'analyse de la sémantique latente

Le principe général de l'analyse de la sémantique latente (*Latent Semantic Analysis*, ou LSA [14, 15]) consiste à définir la signification des mots à partir des contextes dans lesquels ils apparaissent au sein de vastes corpus de textes. Il est en effet possible de déterminer le sens d'un mot à partir de son contexte, dès lors que ce mot est rencontré suffisamment souvent. C'est d'ailleurs de cette manière que nous avons appris une grande partie des mots que nous connaissons : nous les avons rencontrés lors de nos lectures et nous avons petit à petit précisé leur signification, sans que personne ne nous en ait donné la définition. LSA analyse les contextes d'occurrence des mots au sein d'un vaste corpus et réduit le bruit causé par la variabilité de l'emploi de ces mots dans la langue. Chaque mot est représenté par un vecteur dans un espace de plusieurs centaines de dimensions, c'est-à-dire par une suite de centaines de valeurs numériques. Cette représentation vectorielle permet aisément de calculer un vecteur pour une suite de mots, voire un texte entier, en ajoutant simplement les vecteurs des mots qui les composent. Ce formalisme de représentation ne possède pas le côté explicite des représentations symboliques du sens, mais il compense cela par une métrique objective rendant possible des comparaisons de signification entre mots ou groupe de mots, de manière complètement automatique. Par exemple, les deux phrases « J'ai perdu mon chat dans la forêt. » et « Le petit félin a disparu dans les arbres. » sont représentées par deux vecteurs qui sont très proches, indiquant que les significations correspondantes sont voisines. Ce modèle intéresse particulièrement la psycholinguistique, car les performances du modèle s'accordent relativement bien avec celles de sujets humains : les mesures d'association sémantique entre mots calculées par LSA corrèlent de manière satisfaisante avec les jugements produits par les humains [16]. De plus, les courbes d' « apprentissage » du modèle, définies à partir des performances en fonction de la quantité de textes traités peuvent être mise en relation avec celles des enfants qui apprennent la signification des mots à partir de la lecture [14]. LSA constitue donc un solide modèle de représentation sémantique, qui peut être étudié en soi ou servir à construire des modèles cognitifs de plus haut niveau. LSA peut être testé sur le site de l'université du Colorado : http://lsa.colorado.edu

IEA, Intelligent Essay Assessor : évaluer un résumé en le comparant à des copies pré-évaluées

Les étudiants utilisant IEA rédigent un texte correspondant à ce qu'ils connaissent d'un cours et soumettent ce texte à l'ordinateur, qui compare ce texte à des textes-cibles préalablement sélectionnés. Après avoir « entraîné » LSA avec un corpus du domaine (cours), le texte de l'élève est comparé à une ou plusieurs copies-types, sélectionnée(s) par l'enseignant. Deux techniques ont été testées [17], donnant deux types de scores à la copie :

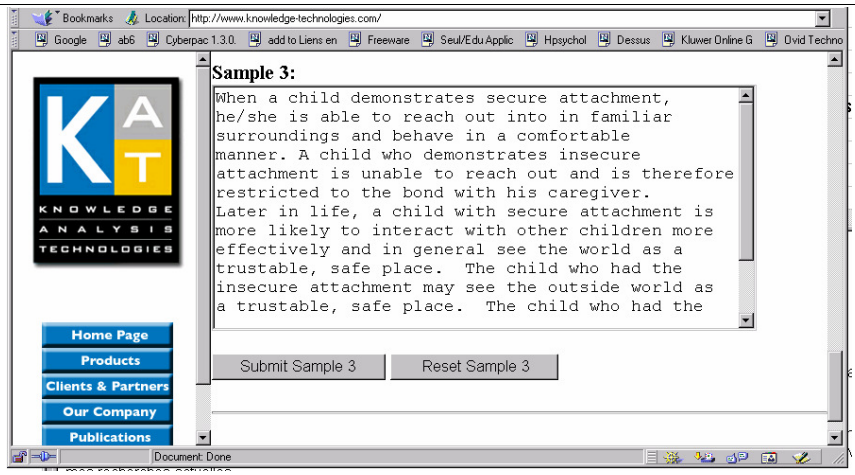
— le score « holistique », qui compare successivement le texte à noter à une série de copies notées au préalable par un jury. La note de la copie sera celle de la série de copies avec laquelle elle entretient la plus grande proximité, calculée par IEA. Une évaluation de ce calcul de score a été faite à partir de 190 copies de biologie, elle montre une corrélation de 0,80 entre les scores des évaluateurs humains et ceux calculés par IEA.

— le score « étalon-or » (*gold standard*), qui compare le texte à noter avec une copie-modèle idéale, réalisée par exemple par l'enseignant. La comparaison peut être globale ou bien

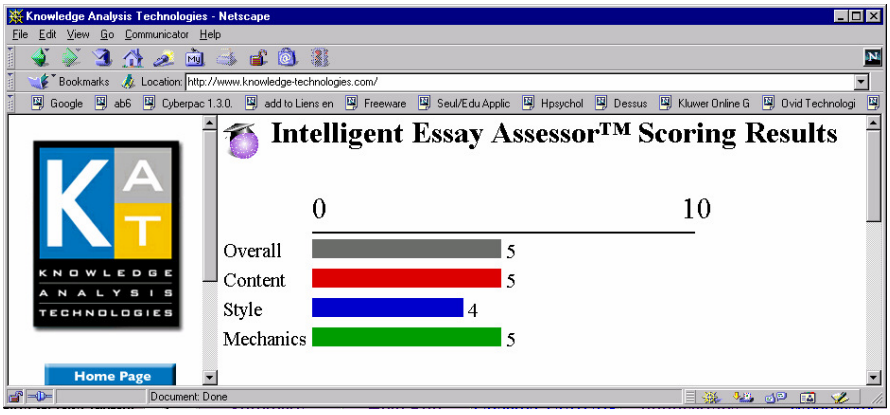
faite paragraphe par paragraphe, de manière à vérifier si l'élève traite correctement chaque notion. Cette méthode est assez proche de la méthode humaine « par impression générale ».

Toutefois, un des avantages majeurs d'IEA, comparé à une évaluation humaine, est que l'étudiant peut soumettre son texte autant de fois que nécessaire, tant que ce dernier n'obtient pas la note voulue. Les auteurs observent d'ailleurs que les notes moyennes des textes soumis à IEA passent de 85/100 — note de la première soumission — à 92/100.

Encadré 3 — Copies d'écran d'*Intelligent Essay Assessor* http://www.knowledge-technologies.com/



Le rédacteur écrit son texte dans la fenêtre ci-dessus. Cliquer sur le bouton « Submit Sample 3 » amène à la fenêtre suivante. Ensuite, IEA évalue le texte par des notes sur dix selon les quatre aspects suivants : *overall* (méthode holistique), *content* (comparaison de type étalon-or), *style* (phrases redondantes), *mechanics* (erreurs d'orthographe et de grammaire).



Apex 1.0 : évaluer le contenu d'un résumé en le comparant au cours

Apex [18, 19, 20] ne compare pas la copie de l'étudiant avec d'autres copies, mais avec le texte d'un cours. Ce dernier est découpé en petites unités par l'enseignant grâce à un système

de balises textuelles, ce qui permettra au logiciel de comparer la copie à certaines parties du cours uniquement. L'étudiant rédige sa copie selon deux objectifs : soit il travaille sur une suite d'unités (un chapitre), soit il répond à une question d'examen. Dans le premier cas, sa copie sera successivement comparée à chaque unité du chapitre, dans le second cas, elle sera appariée aux unités que l'enseignant aura jugées pertinentes pour la question d'examen. Chacune de ces comparaisons correspond en fait à mesurer la proximité des vecteurs correspondant à la copie et à une unité de cours. A chaque fois, *Apex 1.0* indique si la comparaison est bonne, moyenne ou mauvaise. Supposons qu'un étudiant suive un cours de sciences cognitives et traite la question d'examen suivante : « *quelles sont les caractéristiques qui distinguent le connexionnisme des modèles symboliques ?* ». L'enseignant aura par exemple identifié trois unités dans le cours qui doivent être présentes dans la copie de l'étudiant pour constituer une bonne réponse. Le type de retour d'*Apex* pourra alors être le suivant :

- `Unité 42 « Le traitement parallèle » : votre copie couvre bien cette partie (indice=0,79). Bravo !`
- `Unité 45 « Le niveau sub-symbolique » : votre copie couvre très mal cette partie (indice=0,31). Retravaillez votre copie en conséquence.`
- `Unité 40 « Le réseau distribué » : votre copie couvre moyennement cette partie (indice=0,55).`

`Note globale : 11/20`

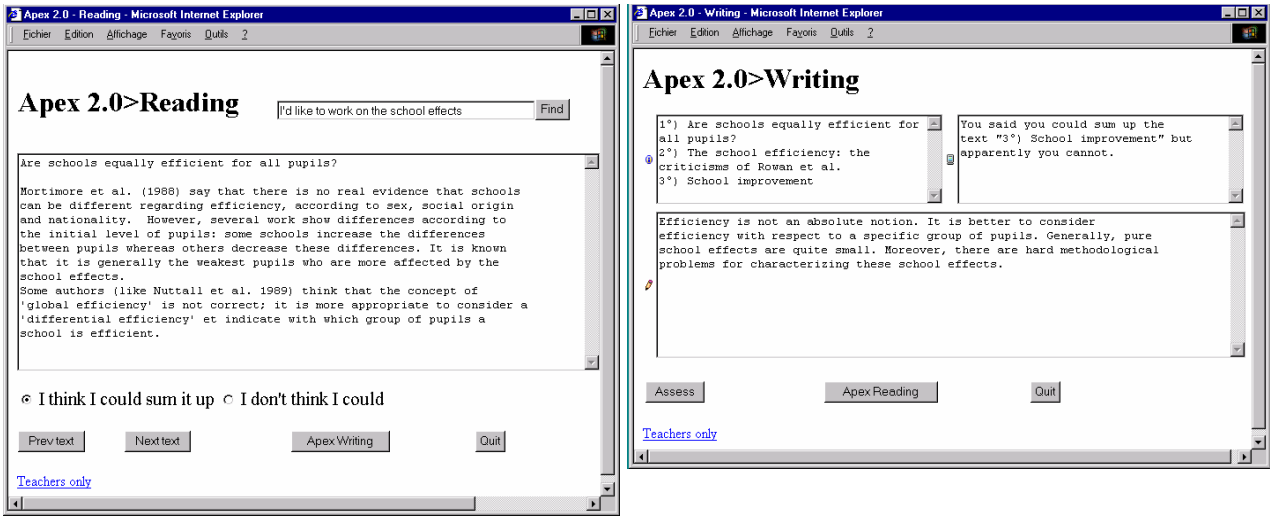
Une note est également calculée, à partir des moyennes des valeurs de comparaison, qui peut être ensuite comparée avec le résultat de corrections humaines. Une corrélation de 0,59 a été obtenue en comparant les copies de 31 étudiants avec les notes qu'avait donné l'enseignant sur un cours de sociologie de l'éducation au niveau licence [20]. Par la suite, cette corrélation a été améliorée (0,62) en procédant préalablement à un découpage de la copie de l'étudiant avant la comparaison [21]. *Apex 1.0* fournit d'autres indications à l'étudiant, notamment concernant la cohérence de sa copie. En effet, la moyenne des similarités sémantiques entre phrases adjacentes est un bon indicateur de la cohérence d'un texte, elle-même ayant un effet sur sa compréhension [15]. En particulier, *Apex 1.0* peut détecter des ruptures de cohérence dans le texte, c'est-à-dire une suite de deux phrases faiblement associées sur le plan sémantique.

Apex 2 : évaluer un résumé pour évaluer la compréhension

Contrairement aux outils précédents qui sont fondés sur la production de l'élève, *Apex 2* [22] s'appuie en plus sur un jugement de compréhension par l'élève de textes lus. L'interaction de

l'élève avec le système comprend deux phases, effectuées en boucle : une phase de lecture de textes, et une phase de production d'un résumé des textes lus. Après chaque lecture de texte, l'élève doit indiquer s'il lui semble possible de résumer le texte. Lorsqu'il le souhaite, l'élève passe à la phase de production : il doit résumer l'ensemble des textes qu'il a estimé pouvoir l'être. C'est là qu'intervient LSA : la comparaison sémantique entre les résumés produits et les textes d'origine permet de montrer à l'élève l'écart entre ce qu'il a cru avoir compris et ce qu'il a vraisemblablement compris. Pour réduire cet écart, l'élève peut de nouveau retravailler son résumé. A tout moment, il peut de nouveau repasser à la phase lecture où de nouveaux textes, là encore judicieusement sélectionnés par LSA en fonction de ce qu'il a déjà lu et de ce qu'il est supposé avoir compris, lui sont proposés. La figure 2 montre un des textes lus par l'étudiant (écran de gauche), le résumé d'un élève (écran de droite, texte en bas), les trois textes qu'il a estimé pouvoir résumer (écran de droite, texte en haut à gauche) et les commentaires d'*Apex 2* à ce sujet (texte en haut à droite).

Figure 2 — Copie d'écran d'*Apex 2*. L'écran de gauche représente la boucle « lecture », dans laquelle l'étudiant prend connaissance du contenu à traiter. L'écran de droite représente la boucle « écriture », dans laquelle l'étudiant peut soumettre à évaluation un résumé de ce qu'il a compris des textes lus.



CONCLUSION

Les principales limites des aides informatiques à la correction de copies sont les suivantes. Premièrement, elles ne prennent pas en compte l'éventuel public-cible de la copie. Même si ce dernier est encore le plus fréquemment l'enseignant, des nouvelles techniques d'évaluation, comme le *portfolio*, montrent que le public-cible peut être l'élève producteur, ou même d'autres personnes hors de la classe. Deuxièmement, les processus de compréhension

d'une copie ne sont pas encore pris en compte : il est difficile de les simuler, car cela nécessite d'avoir une base de connaissances du lecteur, ce qui est difficile à réaliser. Cependant, depuis les premiers logiciels de correction automatique de copies des années 1960, des progrès ont été réalisés [11]. Tout d'abord, une plus grande interactivité des interfaces, une meilleure prise en compte des aspects sémantiques des copies, et enfin une simulation plus fidèle de l'activité humaine de correction.

La recherche actuelle sur la correction automatique de copies se focalise sur les points suivants : correction de réponses courtes [23] et de portfolios, lien entre prise de notes et production d'une copie [24]. De plus, certains aspects de la correction non cités ici, comme l'annotation de la copie par l'enseignant [25], ou encore la détection du plagiat [26] font déjà l'objet de nombreuses recherches. Enfin, l'intégration de tels outils dans de véritables dispositifs d'enseignement à distance fera vraisemblablement l'objet, dans les années à venir, des préoccupations des chercheurs [27]. En effet, s'il est facile de créer autant d'adresses de courrier électronique que d'étudiants, et d'apporter à ces derniers des moyens de consulter ou rechercher des informations, ainsi que de dialoguer, il est beaucoup moins aisé d'évaluer suffisamment rapidement leurs productions, et de leur fournir des conseils appropriés. Les méthodes d'évaluation exposées ci-dessus apporteront une aide importante aux gestionnaires de tels cours à distance.

Repères pour l'action

Quelques conseils à propos du type de questions à formuler, empêchant le plagiat au profit de l'apprentissage [26].

Poser des questions qui nécessitent un traitement de l'étudiant. C'est-à-dire, non pas de trouver des informations, mais de résoudre des problèmes, ou de se poser des questions.

Leur enseigner la « prise de notes verte » et l'éthique de la citation. Dans les cours, faire en sorte que les étudiants distinguent les idées qu'ils ont collectées d'autres de celles qu'ils ont construites en réactions à ces dernières, par exemple, en demandant d'écrire en vert les citations.

Changer les sujets d'examen d'une année à l'autre. Une des sources importantes de plagiat est qu'un étudiant est informé d'un sujet d'examen par la seule lecture du sujet de l'année précédente.

L'enseignant dispose d'une version électronique de la ou des copie(s) à corriger.

Soumettre les copies à divers outils de correction de surface (correcteur orthographique, grammatical, lisibilité).

Éventuellement, en cas de doute, s'assurer que l'étudiant n'a pas plagié. Pour cela, on peut se dire que, si l'étudiant est parvenu facilement à cet extrait, l'enseignant qui corrige la copie pourra le faire encore plus facilement, en tapant dans un moteur de recherche un extrait bien choisi de son mémoire (*i.e*., suffisamment particulier, suffisamment bien écrit, insuffisamment relié au reste du texte).

Utiliser un logiciel d'annotation qui, s'il ne permet pas d'annoter automatiquement la copie, facilite le processus, en proposant des commandes qui insèrent à l'endroit approprié de la copie des commentaires prédéterminés (voir par exemple *HEMP*, *Hypertext Essay Markup Protocol*, à http://www.harpercollege.edu/writ_ctr/hemp/hemp.htm, ou *Question Mark* à http://www.questionmark.com/fra/home.htm

L'enseignant ne dispose pas de la version électronique de la ou des copie(s) à corriger.

Choisir une méthode de correction (*voir ci-dessus*) et s'y tenir pour l'ensemble des copies à corriger.

Quand on corrige une pile de copies, éviter autant que possible les différents effets d'ordre, de contraste, etc. Voir à ce propos les judicieux conseils d'Amigues « La notation des copies en situation d'examen et de classe » disponible à http://recherche.aix-mrs.iufm.fr/publ/voc/n1/amigues4/index.html (accédé le 20 juin 2003).

Adresse web des différents outils de correction automatique.

Intelligent Essay Assessor : http://www.knowledge-technologies.com/IEA.html

Summary Street : http://lsa.colorado.edu/summarystreet/

POUR EN SAVOIR PLUS…

Dessus, P., & Lemaire, B. (1999). Apex, un système d'aide à la préparation des examens. *Sciences et Techniques Educatives, 6*(2), 409-417. [Une description plus complète d'*Apex*, en français.]

Legros, D., & Crinon, J. (2002). *Psychologie des apprentissages et multimédia*. Paris : Colin. [Un des meilleurs ouvrages récents et en français sur les effets des différents outils informatiques sur l'apprentissage. Lire en particulier le chapitre 5, « Apprendre à écrire »]

Noizet, G., & Caverni, J.-P. (1978). *Psychologie de l'évaluation scolaire*. Paris: P.U.F. [Grand classique sur les aspects psychologiques de l'évaluation.]

REMERCIEMENTS

Nous remercions vivement Françoise Campanale, Dany Hecquet, Patrick Mendelsohn et Laurent Tarillon pour leurs commentaires d'une version précédente de ce chapitre.

REFERENCES

[1] Dessus, P. (2001). Aides informatisées à la production d'écrits, une revue de la littérature. *Sciences et Techniques Éducatives, 8*(3-4), 413-433.

[2] Kukich, K. (2000). Beyond Automated Essay Scoring. *IEEE Intelligent Systems, 15*(5), 22-27.

[3] French, P. (1998). *Developments in the provision of quality electronic summative assessments* (Rapport de recherche). Lower Hutt : The Open Polytechnic of New Zealand.

[4] Schriver, K. A. (1989). Evaluating text quality: The continuum from text-focused to reader-focused methods. *IEEE Trans. on Professional Communication*, *32*(4), 238-255.

[5] Hayes, J. R. (1998). Un nouveau cadre pour intégrer cognition et affect dans la rédaction. In A. Piolat & A. Pélissier (Eds.), *La rédaction de textes, approche cognitive* (pp. 51-101). Lausanne : Delachaux et Niestlé.

[6] Hamp-Lyons, L. (1992). Holistic writing assessment for LEP students. *Proc. Second National Research Symposium on Limited English Proficient Student Issues*. Washington.

[7] Noizet, G., & Caverni, J.-P. (1978). *Psychologie de l'évaluation scolaire*. Paris : P.U.F.

[8] Huteau, M. (1996). L'évaluation par les notes et par les tests. In A. Lieury (Ed.), *Manuel de psychologie de l'éducation et de la formation* (pp. 271-302). Paris : Dunod.

[9] Miller, T. (à paraître). Essay assessment with Latent Semantic Analysis. *Journal of Educational Computing Research*.

[10] Dale, E. & Chall, J. S. (1948). A formula for predicting readability. *Educational Research Bulletin*, *27*, 37-53.

[11] Page, E. B. (1994). New computer grading of student prose. *Journal of Experimental Education, 62*(2), 127-142.

[12] Burstein, J., Kukich, K., Wolff, S., Lu, C., & Chodorow, M. (1998). Computer analysis of essays. *Communication au NCME Symposium on Automated Scoring*, Montréal.

[13] Christie, J. R. (1999). Automated Essay Marking—for both style and content. *Proc. third Computer Assisted Assessment Conference*. Loughborough.

[14] Landauer, T. K., Foltz, P. W., & Laham, D. (1998). An introduction to Latent Semantic Analysis. *Discourse Processes, 25*(2/3), 259-284.

[15] Lemaire, B., & Dessus, P. (2003). Modèles cognitifs issus de l'Analyse de la sémantique latente. *Cahiers Romans de Sciences Cognitives, 1*(1), 55-74.

[16] Foltz, P. W. (1996). Latent semantic analysis for text-based research. *Behavior Research Methods, Instruments and Computers, 28*(2), 197-202.

[17] Foltz, P. W., Laham, D., & Landauer, T. K. (1999). Automated essay scoring : applications to Educational Technology. *Proc. Conf. ED-MEDIA'99*. Seattle.

[18] Dessus, P., & Lemaire, B. (1999). Apex, un système d'aide à la préparation des examens. *Sciences et Techniques Educatives, 6*(2), 409-417.

[19] Dessus, P., Lemaire, B., & Vernier, A. (2000). Free-text assessment in a Virtual Campus. In K. Zreik (Ed.), *Proc. International Conference on Human System Learning (CAPS'3)* (pp. 61-76). Paris : Europia.

[20] Lemaire, B., & Dessus, P. (2001). A system to assess the Semantic Content of Student Essays. *Journal of Educational Computing Research, 24*(3), 305-320.

[21] Gounon, P., & Lemaire, B. (2002). Semantic comparison of texts for learning environments. In F. J. Garijo, J. C. R. Santos, & M. Toro (Eds.), *Advances in Artificial Intelligence (IBERAMIA 2002)* (pp. 724-733). Berlin : Springer.

[22] Dessus, P., & Lemaire, B. (2002). Using production to assess learning: An ILE that fosters Self-Regulated Learning. In S. A. Cerri, G. Gouardères, & F. Paraguaçu (Eds.), *Intelligent Tutoring Systems (ITS 2002)* (pp. 772-781). Berlin : Springer.

[23] Hirschman, L., Breck, E., Light, M., Burger, J. D., & Ferro, L. (2000). The Intelligent Essay Assessor. *IEEE Intelligent Systems, 15*(5), 31-35.

[24] Zaoui, S. (2003). Effets de l'évaluation des connaissances sur la prise de notes et sur l'apprentissage dans un environnement informatique. *9e Journée d'Étude sur le Traitement Cognitif des Systèmes d'Information Complexes (JETCSIC)*. Dijon : Université de Bourgogne, LEAD.

[25] Stephens, D., Sargent, G., & Brew, I. (2001). Comparison of assessed work marking software: implications for the ideal Integrated Marking Tool (IMT). *Proc. 5th International CAA Conference.* Loughborough.

[26] McKenzie, J. (1998). The new plagiarism. *From Now On, 7*(8).

[27] Streeter, L., Psotka, J., Laham, L., & MacCuish, D. (2002). The Credible Grading Machine: Automated Essay Scoring in the DoD. *Conf. « Interservice/Industry, Simulation and Education »* (I/ITSEC). Orlando.