



## Discours

Revue de linguistique, psycholinguistique et informatique. A journal of linguistics, psycholinguistics and computational linguistics

11 | 2012  
Varia

---

# Évaluation automatique de textes et cohésion lexicale

Yves Bestgen

---



### Édition électronique

URL : <http://journals.openedition.org/discours/8724>

DOI : 10.4000/discours.8724

ISSN : 1963-1723

### Éditeur :

Laboratoire LATTICE, Presses universitaires de Caen

### Référence électronique

Yves Bestgen, « Évaluation automatique de textes et cohésion lexicale », *Discours* [En ligne], 11 | 2012, mis en ligne le 23 décembre 2012, consulté le 30 avril 2019. URL : <http://journals.openedition.org/discours/8724> ; DOI : 10.4000/discours.8724

---



*Discours* est mis à disposition selon les termes de la licence Creative Commons Attribution - Pas d'Utilisation Commerciale - Pas de Modification 4.0 International.





Revue de linguistique, psycholinguistique et informatique

<http://discours.revues.org/>

## Évaluation automatique de textes et cohésion lexicale

---

Yves Betsgen

Centre for English Corpus Linguistics  
Université catholique de Louvain

.....  
Yves Bestgen, «Évaluation automatique de textes et cohésion lexicale», *Discours* [En ligne], 11 | 2012, mis en ligne le 23 décembre 2012.

.....  
URL: <http://discours.revues.org/8724>

.....  
Titre du numéro: *Varia*  
Coordination: Olivier Ferret et Nicolas Hernandez

**revues.org**  
CENTRE POUR L'ÉDITION ÉLECTRONIQUE OUVERTE  
CENTRE FOR OPEN ELECTRONIC PUBLISHING

 discours

 Presses  
universitaires  
de Caen



# Évaluation automatique de textes et cohésion lexicale

---

Yves Bestgen

Centre for English Corpus Linguistics  
Université catholique de Louvain

.....

L'évaluation automatique de textes connaît actuellement un succès grandissant en raison de son importance dans le champ de l'éducation et, tout particulièrement, de l'apprentissage des langues étrangères. Si des systèmes efficaces ont été développés ces quinze dernières années, peu d'entre eux prennent en compte le niveau discursif. Récemment, quelques recherches ont proposé de remédier à cette lacune au moyen de mesures automatiques de la cohésion lexicale obtenues à partir d'une analyse sémantique latente, mais les résultats n'ont pas été conformes aux attentes. En s'inspirant d'un modèle bien connu de l'expertise rédactionnelle, la présente recherche propose d'employer un nouvel indice de cohésion dérivé des travaux en segmentation thématique de textes. L'efficacité de cet index est confirmée au travers de l'analyse d'un corpus de 223 textes d'apprenants de l'anglais comme langue étrangère. La conclusion discute les limitations principales de cette étude exploratoire et propose des pistes de développement.

**Mots clés :** évaluation automatique de textes, cohésion lexicale, analyse sémantique latente, segmentation thématique, stratégies rédactionnelles, diversité lexicale

.....

*Automatic essay grading is currently experiencing a growing popularity because of its importance in the field of education and, particularly, in foreign language learning. While several efficient systems have been developed over the last fifteen years, almost none of them take the discourse level into account. Recently, a few studies proposed to fill this gap by means of automatic indexes of lexical cohesion obtained from Latent Semantic Analysis, but the results were disappointing. Based on a well-known model of writing expertise, the present study proposes a new index of cohesion derived from work on the thematic segmentation of texts. The efficiency of this index is supported through the analysis of a corpus of 223 essays of learners of English as a foreign language. The conclusion discusses the main limitations of this exploratory study and proposes further avenues for development.*

**Keywords :** automatic essay grading, lexical cohesion, Latent Semantic Analysis, thematic segmentation, writing strategies, lexical diversity

## 1. Introduction

- 1 Répondre à une question ouverte lors d'un examen ou rédiger une dissertation sont des dispositifs évaluatifs, mais également des outils de formation particulièrement importants dans le champ de l'éducation. Non seulement ils permettent une évaluation plus fine et plus riche des compétences de l'étudiant qu'un questionnaire à réponses fermées (Magliano et Graesser, 2012), mais surtout la mise en évidence par l'évaluation des points forts et des points faibles de ces textes est formative pour l'étudiant. La lourdeur de l'évaluation et les fréquents désaccords observés lorsque les avis de plusieurs évaluateurs sont comparés limitent souvent le recours à ces modalités d'évaluation (Miller, 2003).

- 2 Dans ce contexte, l'évaluation automatique de textes est devenue un enjeu capital. Selon la tâche demandée à l'étudiant, deux grands types d'approches ont été développés. Lorsqu'il s'agit de faire preuve d'un niveau suffisant de connaissance sur un sujet donné, l'approche classique consiste à comparer le texte de l'étudiant avec la réponse attendue, le plus souvent en évaluant son contenu sémantique (Rehder *et al.*, 1998; Landauer et Psozka, 2000). Lorsque le contenu du texte à produire est plus libre, comme dans le cas de dissertations, on emploie des indices linguistiques plus ou moins fortement en corrélation avec la qualité de l'essai. Le prototype de ce genre d'approche est certainement *e-Rater*, développé par *Educational Testing Services* (ETS), qui se base, entre autres, sur les erreurs d'orthographe ou grammaticales qui peuvent être détectées automatiquement, la richesse lexicale et la longueur de l'essai (Burstein, Chodorow et Leacock, 2004; Burstein, 2009). De tels systèmes sont particulièrement utiles dans le domaine de l'apprentissage d'une langue étrangère où être capable de rédiger un texte de qualité est un des objectifs principaux de la formation.
- 3 Le développement de procédures automatiques d'évaluation requiert la collaboration de chercheurs en psychologie et éducation, en linguistique et en traitement automatique du langage (TAL). La psychologie et les sciences de l'éducation apportent des modèles de la production de textes et des études de l'évolution des compétences tout au long du développement cognitif. La linguistique propose des indices susceptibles de mesurer la qualité des textes. Le TAL développe des techniques qui permettent d'extraire automatiquement ces indices des textes.
- 4 Si ce champ de recherche appliquée a déjà une longue histoire, les premiers systèmes datant des années soixante (Page, 1966), la prise en compte de la dimension discursive n'est encore qu'à ses balbutiements (Deane et Quinlan, 2010). Récemment, une série d'études ont essayé de faire profiter l'évaluation automatique des textes des résultats récents en linguistique, en psychologie et en TAL qui permettent la prise en compte de la cohésion dans les textes. D'une manière générale, les résultats obtenus sont décevants et même contre-intuitifs : les meilleurs textes sont ceux dont la cohésion, mesurée de manière automatique, est la plus faible. Après avoir présenté ces travaux et leurs principales limitations, la présente recherche propose d'employer un nouvel indice de cohésion dérivé des travaux en segmentation thématique de textes (Morris et Hirst, 1991; Choi, 2000). L'efficacité de cette approche est évaluée au travers de l'analyse d'un corpus de 223 textes d'apprenants de l'anglais comme langue étrangère. Le résultat le plus important de cette étude exploratoire est de souligner la nécessité d'une approche pluridisciplinaire du discours lorsqu'on vise des objectifs appliqués.

## 2. Mesurer automatiquement la cohésion lexicale grâce à l'analyse sémantique latente

- 5 Si quelques recherches ont proposé de s'appuyer sur les chaînes de référence pour mesurer la cohésion dans des textes (Miltakaki et Kukich, 2004; Lapata et Barzilay, 2005), l'approche la plus fréquemment employée pour atteindre cet objectif est sans

aucun doute l'analyse sémantique latente (*Latent Semantic Analysis* – LSA). Au dire même de Landauer *et al.* (2007: 10), éditeurs du *Handbook of Latent Semantic Analysis*, cette technique «*accurately estimates passage coherence*». La présente section décrit les caractéristiques les plus importantes de cette technique et sert donc d'introduction à la section suivante qui discute les résultats obtenus lorsqu'elle est appliquée à l'analyse automatique de la cohésion de textes dans le but de prédire leur qualité. Une présentation plus technique est donnée dans Landauer, Foltz et Laham (1998).

- 6 Depuis une vingtaine d'années, l'analyse sémantique latente (désormais ASL) est utilisée dans un nombre de plus en plus important de recherches en TAL, psychologie, et sciences de l'éducation (Landauer *et al.*, 2007). L'ASL vise à construire un espace sémantique de très grandes dimensions à partir de l'analyse statistique des cooccurrences dans un corpus de textes. Cet espace sémantique est ensuite employé pour estimer la similarité sémantique entre des mots, des phrases, des paragraphes et même des textes. L'idée sous-jacente, bien établie en linguistique, est que des mots similaires apparaissent dans des contextes similaires. Le point de départ de l'analyse est un tableau lexical qui contient le nombre d'occurrences de chaque mot dans chacun des documents qui composent un corpus souvent constitué de plusieurs millions de mots, un document pouvant être un texte, un paragraphe ou même une phrase. Pour extraire les dimensions sémantiques, ce tableau de fréquence fait l'objet d'une décomposition en valeurs singulières, une sorte d'analyse factorielle, qui en extrait les dimensions orthogonales les plus importantes. Cette compression de la matrice de cooccurrences initiale apporte deux bénéfices : elle élimine les cooccurrences les plus anecdotiques et elle permet la prise en compte, non seulement des relations de premier ordre, celles qui établissent comme similaires des mots qui apparaissent dans un même contexte, mais aussi des relations de deuxième ordre pour lesquelles deux mots sont similaires même s'ils n'apparaissent pas dans un même contexte, parce que les contextes dans lesquels ils apparaissent sont similaires.
- 7 Tant les mots que les segments originaux sont positionnés dans cet espace sémantique, ce qui permet de mesurer leur proximité. Plus précisément, le sens de chaque mot y est représenté par un vecteur. Pour mesurer la similarité sémantique entre deux mots, on calcule le cosinus entre les vecteurs qui les représentent. Plus deux mots sont sémantiquement proches, plus les deux vecteurs qui les représentent pointent dans la même direction et donc plus leur cosinus se rapproche de 1. À titre d'illustration, le Tableau 1 présente quelques-uns des plus proches voisins des mots *amour* et *vol*, calculés à partir d'une base sémantique dérivée de l'analyse de 206 contes de Maupassant (voir Bestgen et Cabiliaux, 2002). Le caractère ambigu du sens de *vol* apparaît très nettement dans les deux types de voisins qui lui sont associés.
- 8 Les mêmes calculs peuvent être effectués sur les vecteurs qui représentent les documents analysés. La technique est encore plus générale puisqu'elle permet de calculer le vecteur qui correspond à un groupe de mots même si ce groupe de mots ne constitue pas un document analysé en tant que tel. Il est ainsi possible d'analyser

la proximité sémantique entre deux paragraphes ou deux phrases qui se suivent dans un texte, même si ce texte n'était pas présent dans le corpus dont a été extrait l'espace sémantique (Foltz, Kintsch et Landauer, 1998).

Mot	<i>Amour</i> Cosinus	Mot	<i>Vol</i> Cosinus
passion	0,70	malfaiteur	0,36
tendresse	0,66	gendarme	0,36
aimer	0,58	tourbillonner	0,32
passionné	0,56	délit	0,32
caresse	0,55	corbeau	0,31
cœur	0,54	justice	0,31

Tableau 1. Les plus proches voisins d'*amour* et de *vol* selon la mesure du cosinus

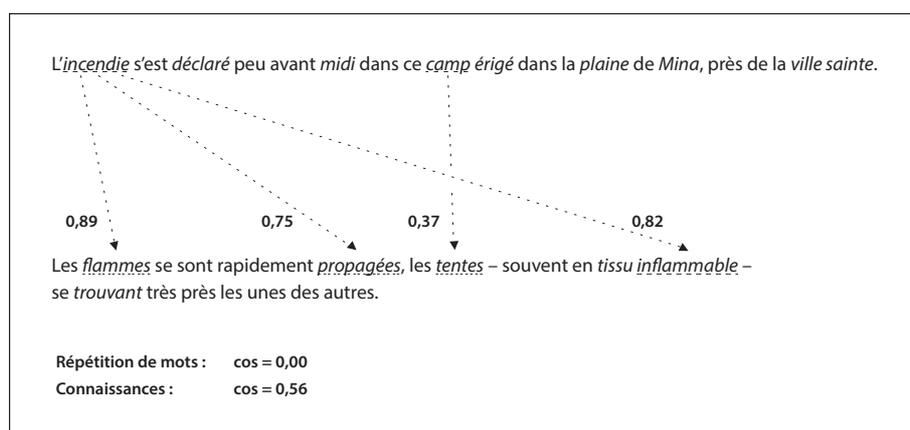


Figure 1. Cohésion lexicale estimée sur la base de la simple répétition de mots ou d'une ASL pour deux phrases contiguës d'un article du journal *Le Monde*

- 9 L'intérêt majeur de l'ASL pour la mesure de la cohésion lexicale est qu'elle apporte une réponse aux problèmes posés par les phrases qui relèvent d'un même thème tout en ne partageant aucun mot commun ou par la présence de synonymes et d'hyperonymes<sup>1</sup>. En cela, l'ASL permet d'aller au-delà du modèle classique vectoriel (Manning et Schütze, 1999 : 539 *sq.*) fréquemment employé pour mesurer la cohésion lexicale entre des phrases (Hearst, 1997; Choi, 2000; Kehagias, Pavlina

1. L'ASL n'est pas la seule technique proposée pour répondre à ces problèmes (voir, par exemple, Morris et Hirst, 1991; Kozima, 1993; Ferret, 2002).

et Petridis, 2003). Dans le modèle vectoriel, la similarité entre deux phrases est basée uniquement sur les mots communs. Ainsi, comme le montre la Figure 1, des phrases thématiquement très similaires peuvent être déclarées non liées parce qu'elles ne partagent aucun mot. L'ASL, en revanche, met en évidence une similarité importante parce qu'elle prend en compte le fait que des mots comme *incendie* et *flamme* ( $\cos = 0,89$ ) ou comme *camp* et *tente* ( $\cos = 0,37$ ) sont liés.

10 Cette figure illustre également la principale manière dont l'ASL est employée pour mesurer la cohésion lexicale dans un texte : calculer le cosinus entre toutes les paires de phrases contiguës<sup>2</sup> (Foltz, Kintsch et Landauer, 1998; Bestgen, Degand et Spooren, 2006; Foltz, 2007). Le degré de cohésion du texte dans son ensemble est alors estimé par la moyenne de tous ces cosinus.

11 Afin de montrer que l'ASL permet une quantification efficace de la cohésion, une série de recherches ont eu recours à des textes dont le degré de cohésion avait été manipulé dans le but de permettre une étude des facteurs qui affectent les processus cognitifs à l'œuvre lors de la lecture (Foltz, Kintsch et Landauer, 1998; Foltz, 2007; Graesser et McNamara, 2011). Tout particulièrement, ces travaux ont comparé le degré de cohésion lexicale estimé par l'ASL pour deux versions des mêmes textes : la version originale et une version réécrite dans le but d'en accroître l'intelligibilité. Ces analyses ont systématiquement mis en évidence une cohésion, mesurée par l'ASL, plus grande dans les textes réécrits.

### 3. ASL, cohésion et évaluation automatique de textes

12 Ces trois dernières années, quelques recherches ont tenté de déterminer s'il était possible de prédire la qualité de textes produits en langue maternelle et en langue seconde à partir d'indices de cohésion lexicale issus de l'ASL (Bestgen, Lories et Thewissen, 2010; Crossley, Roscoe et McNamara, 2011; Crossley et McNamara, 2012). Dans toutes ces études, la procédure employée consiste à calculer la corrélation entre ces indices de cohésion et les évaluations de la qualité de textes par des experts. Les résultats sont sans appel : soit aucune relation n'est observée, soit la relation est l'inverse de celle qui était attendue, les meilleurs textes pour les experts humains étant les moins cohésifs selon la procédure automatique (Bestgen, Lories et Thewissen, 2010; Crossley et McNamara, 2012).

13 Comment expliquer ces résultats ? Il semble difficile d'incriminer les indices linguistiques de la cohésion pris en compte, ceux-ci étant justifiés par de très nombreuses recherches en linguistique depuis les travaux séminaux de Halliday et

---

2. Le cosinus peut également être calculé entre les paragraphes contigus ou entre chaque phrase d'un paragraphe et le paragraphe lui-même ou encore entre chaque phrase d'un texte et celui-ci (Foltz, 2007). Pour prédire la qualité d'un texte, c'est toutefois le cosinus moyen entre les phrases contiguës qui est le plus souvent employé (Crossley et McNamara, 2012). De plus, dans les textes qui sont analysés dans la présente étude, la segmentation en paragraphes n'est pas fiable parce que certains rédacteurs ont introduit un alinéa après chaque phrase.

Hasan (1976). De même, l'implémentation de ces indices, même si elle ne peut pas être considérée comme parfaite, repose sur des techniques de TAL qui ont fait leur preuve (Bestgen, 2006; Landauer *et al.*, 2007). De plus, comme indiqué ci-dessus, plusieurs études ont montré que l'ASL est efficace pour distinguer les textes ou extraits de textes très et peu cohésifs (Foltz, Kintsch et Landauer, 1998; Bestgen, Degand et Spooen, 2006; Graesser et McNamara, 2011).

14 En revanche, on peut s'interroger à propos du postulat sur lequel reposent ces analyses: le lien entre cohésion forte et haute qualité des textes (Crossley et McNamara, 2012). Un bon texte doit-il nécessairement présenter une plus forte cohésion lexicale entre les phrases contiguës? On peut en douter comme le souligne Foltz (2007: 177): «*better essays often tend to have low coherence*», même s'il n'apporte pas d'argument empirique à l'appui de cette assertion. Ce n'est pas tant la quantité que la qualité qui compte. En d'autres mots, ce n'est pas la force de la cohésion lexicale mesurée par le cosinus moyen entre les phrases contiguës qui importe, mais la distribution de cette cohésion dans des textes de bonne et de moins bonne qualité. De plus, on peut penser que le cosinus moyen interphrases est influencé par d'autres facteurs qui interviennent dans la qualité d'un texte, comme la richesse lexicale: moins le vocabulaire d'un texte est riche et plus ce cosinus est élevé puisque rien n'est lexicalement plus semblable que deux mots identiques.

15 Afin de proposer une hypothèse psychologiquement crédible qui lie cohésion et qualité des textes, la section suivante synthétise des travaux menés dans les années quatre-vingt par des chercheurs dont l'objectif était de modéliser le développement des compétences rédactionnelles d'enfants et d'adolescents.

#### 4. Expertise, stratégies rédactionnelles et cohésion lexicale

16 S'appuyant sur une analyse détaillée de protocoles verbaux recueillis lors de la rédaction de textes produits par des enfants, des adolescents et des adultes, Scardamalia et Bereiter (1987) ont identifié deux stratégies rédactionnelles, la stratégie des connaissances rapportées ou *knowledge-telling strategy* et celle des connaissances transformées ou *knowledge-transforming strategy* (Chanquoy et Alamargot, 2003). La première est la seule à laquelle les rédacteurs novices peuvent avoir recours. Elle consiste en une série de cycles de récupération d'informations en mémoire et de formulation de celles-ci (McCutchen, 1996). À partir du sujet du texte, le rédacteur cherche une idée pertinente, il retranscrit directement les informations que cette idée permet de récupérer et ensuite cherche une autre idée pouvant être produite. Comme le soulignent Chanquoy et Alamargot (2003), cette stratégie permet aux rédacteurs novices de produire un texte cohérent malgré l'importance de la charge cognitive requise par cette tâche. Dans ce texte, les informations sont organisées exclusivement sur les bases de leur association en mémoire à long terme (Galbraith, 2009). Aucune réorganisation n'est effectuée.

17 La stratégie des connaissances transformées complète la stratégie précédente en ajoutant à la récupération des informations des processus de transformation et

d'organisation de celles-ci. La rédaction d'un texte est traitée comme un problème à résoudre qui requiert planification et prise en compte de l'audience à laquelle le texte est destiné. La dimension pragmatique et rhétorique de la communication devient centrale. Les informations récupérées en mémoire sont approfondies et restructurées afin de mieux répondre aux objectifs communicationnels du rédacteur (Chanquoy et Alamargot, 2003).

18 Comme l'ont souligné plusieurs chercheurs (McCutchen, 1996; Chanquoy et Alamargot, 2003; Galbraith, 2009), la stratégie des connaissances transformées ne peut être mise en pratique que si le rédacteur dispose de ressources cognitives suffisantes et donc si une partie importante des processus à mettre en œuvre lors de la rédaction d'un texte (recherche en mémoire, retranscription, gestion de l'orthographe) est automatisée, ce qui n'est pas le cas pour le rédacteur novice, ni même pour le rédacteur expert lorsque la situation de production est coûteuse en ressources cognitives.

19 Dans le cadre de ce modèle du développement des capacités rédactionnelles, les textes de moindre qualité devraient se présenter comme une suite non structurée de blocs de phrases fortement connectées entre elles, mais les relations entre ces blocs résulteraient seulement de leurs connexions au sujet du texte alors que la structure des textes de meilleure qualité serait beaucoup plus diffuse, basée sur des relations plus complexes. Cette analyse a deux implications pour l'étude de la relation entre cohésion et qualité des textes.

- Tout d'abord, elle plaide pour l'emploi de techniques qui permettent d'analyser la structure interne des textes et non en faveur du calcul d'un score global de cohésion par texte. Tout particulièrement, il est nécessaire de disposer de techniques de TAL capables de mettre en évidence la structure en blocs isolés composés de phrases fortement connectées, postulée pour les textes de moindre qualité. De telles techniques ont été développées dans le champ de la segmentation thématique de textes sur la base de la cohésion lexicale (Morris et Hirst, 1991; Kozima, 1993; Choi, Wiemer-Hastings et Moore, 2001). Parmi ces techniques, la procédure proposée par Choi (2000) est fréquemment considérée comme état de l'art et employée comme point de repère pour évaluer d'autres techniques. Choi, Wiemer-Hastings et Moore (2001) ont proposé de combiner cette procédure avec l'ASL et Bestgen (2006) a confirmé qu'une telle combinaison en améliore l'efficacité.
- La deuxième implication est qu'il est souhaitable d'analyser des textes produits dans une situation susceptible de faire apparaître les deux types de stratégie. Galbraith (2009: 12-13) a récemment souligné que la production de textes en langue étrangère remplit parfaitement cette condition: *«L2 language proficiency would be expected to affect not just how well-formed the written product is from a linguistic point of view, but also the writer's capacity to engage in the higher level problem-solving activities characteristic of expert writing. Thus, even when L2 production is linguistically accurate, to the extent*

*that L2 language production in L2 remains more effortful than in L1 one might expect writers to be less able to engage in goal directed creation of content and the quality of the text to suffer accordingly» (voir aussi Piérard et Bestgen [2008] pour le même argument en production orale).*

- 20 L'objectif de la présente recherche est d'apporter une première vérification empirique de ces conjectures en analysant 223 essais produits en langue étrangère et en comparant l'efficacité du cosinus interphrases et de la segmentation automatique de textes pour prédire les évaluations de la qualité des textes effectuées par deux juges.

## 5. Méthode

### 5.1. Corpus de textes

- 21 Les essais analysés dans cette étude ont été extraits de la seconde version de l'*International Corpus of Learner English* (ICLE) (Granger *et al.*, 2009), un corpus d'essais écrits par des apprenants de l'anglais comme langue étrangère d'un niveau intermédiaire à un niveau avancé. ICLE est le résultat de dix années de collaboration entre un grand nombre d'universités internationales. Il contient plus de deux millions de mots produits par 3 640 étudiants de 11 langues maternelles différentes. Au total, 223 essais argumentatifs ont été sélectionnés d'une manière aléatoire dans trois sous-corpus ICLE : 74 essais du composant français, 71 du composant allemand et 78 du composant espagnol. Pour pouvoir être choisis, les essais devaient répondre à deux critères : ils devaient compter entre 500 et 900 mots et être argumentatifs. Ils ont été choisis dans le cadre d'un projet de doctorat mené au Centre for English Corpus Linguistics, Louvain-la-Neuve, Belgique (Thewissen, 2012).

### 5.2. Évaluations des textes par des experts

- 22 Deux évaluateurs professionnels, ayant participé à l'évaluation de la qualité de textes pour l'université de Cambridge, ont été invités à évaluer chacun des 223 textes. On leur a notamment demandé d'attribuer un score holistique à chaque texte afin de donner leur impression globale de la qualité de celui-ci sur la base des descripteurs du Cadre européen commun de la référence pour les langues (CECR) (Council of Europe, 2001). Le CECR inclut six niveaux de compétence qui peuvent être décomposés en trois groupes de deux, à savoir A1 et A2 (niveau de compétence élémentaire), B1 et B2 (niveau de compétence intermédiaire), C1 et C2 (niveau de compétence avancé). Les juges ne pouvaient pas employer les niveaux A1 et A2 parce que le CECR spécifie que les étudiants ont besoin d'être au moins au niveau de compétence B1 pour pouvoir entreprendre la rédaction d'un essai. Les évaluateurs pouvaient employer un signe + ou un signe – pour spécifier plus finement la qualité des textes dans chaque niveau de compétence (par exemple B2– ou C1+). Les évaluations obtenues ont été transformées en nombres sur une échelle numérique de 11 points (c.-à-d. B1–=0,66, B1=1, B1+=1,33, etc. jusqu'à C2=4). La corrélation

entre les deux évaluateurs sur cette échelle de notation est de 0,69. Afin d'attribuer un score CECR final à chaque texte, nous avons calculé les moyennes des scores donnés par les deux évaluateurs.

### 5.3. Analyse automatique des essais

#### 5.3.1. Cohésion lexicale estimée par ASL

23 Pour estimer les proximités entre les paires de phrases de chaque texte, un espace sémantique a été extrait du corpus *TASA – General Reading up to 1st Year College*, qui est le corpus de référence pour les travaux psycholinguistiques qui s'appuient sur l'ASL (Landauer, Foltz et Laham, 1998; Landauer *et al.*, 2007). Ce corpus est composé d'extraits de textes, d'une longueur moyenne approximative de 290 mots, obtenus par un échantillonnage aléatoire de textes que lisent les élèves et les étudiants américains. La version à laquelle T.K. Landauer (Institute of Cognitive Science, University of Colorado, Boulder) nous a donné accès contient 44 486 documents et approximativement 12 millions de mots<sup>3</sup>. Ce corpus a été lemmatisé au moyen de *TreeTagger* (Schmid, 2003) et une série de mots outils (*and, be, the, that...*) ont été supprimés ainsi que tous les mots dont la fréquence totale est inférieure à 2. La matrice de cooccurrences a été soumise à une décomposition en valeurs singulières au moyen du programme SVDPACKC (Berry *et al.*, 1993) et les 300 premiers vecteurs propres ont été conservés, ce nombre étant habituellement considéré comme optimal (Landauer, Laham et Derr, 2004).

24 Cet espace sémantique a été utilisé pour calculer les cosinus entre les phrases adjacentes de chaque texte et la moyenne de ces cosinus est utilisée comme estimation de la cohésion lexicale de chacun des textes.

#### 5.3.2. Longueur moyenne des segments

25 La procédure proposée par Choi (2000; Choi, Wiemer-Hastings et Moore, 2001) pour segmenter des textes a été employée. Elle est composée de trois étapes. Tout d'abord, le document à segmenter est découpé en unités textuelles minimales, dans le cas présent il s'agit des phrases. Les mots composant ces phrases sont soumis à différents traitements, comme la suppression de mots peu informatifs sur le thème du texte (article, pronom, verbes très fréquents...) et une lemmatisation. Ensuite, une mesure de similarité entre toutes les paires de phrases prises deux à deux est calculée. Enfin, le document est segmenté de façon récursive en fonction des frontières entre les unités textuelles qui maximisent la somme des similarités moyennes à l'intérieur des segments ainsi constitués. Le nombre de segments à constituer dans un texte donné est déterminé automatiquement par l'algorithme. L'ASL intervient lors de la deuxième étape, les similarités entre les phrases étant obtenues en appliquant la métrique du cosinus aux vecteurs pondérés par les dimensions sémantiques dérivées

3. Un espace sémantique extrait d'une version légèrement différente de ce corpus est disponible sur le site : <http://LSA.colorado.edu>.

d'un corpus. Choi (2000) et Choi, Wiemer-Hastings et Moore (2001) ont montré que cet algorithme était plus efficace qu'une série d'autres approches telles que *TextTiling* de Hearst, *DotPlot* de Reynar, *Segmenter* de Kan *et al.* et le *Maximum-Probability Segmentation Algorithm* de Utiyama et Isahara.

- 26 L'indice extrait de cette analyse est la longueur moyenne d'un segment dans un texte qui présente l'avantage sur le nombre de segments de prendre en compte la longueur du texte (2 segments dans un texte de 20 phrases correspondent à 4 segments dans un texte de 40 phrases).

## 6. Résultats

- 27 Le Tableau 2 présente les moyennes et écarts types des deux indices calculés ainsi que la corrélation entre ceux-ci et l'évaluation des experts. Les deux corrélations sont significatives pour un seuil de 0,0001 ( $N=223$ ).

Indices	Moyenne	Écart type	Corrélation avec l'évaluation
Cosinus moyen entre les phrases contiguës	0,18	0,07	-0,31
Longueur moyenne des segments	18,29	10,06	0,26

Tableau 2. Indices de cohésion : statistiques descriptives et corrélation avec l'évaluation

- 28 La corrélation obtenue pour l'indice de cohésion lexicale extrait de l'ASL confirme les observations récentes : plus un essai présente un score de cohésion élevé, moins il est évalué positivement par les experts. Cette relation négative est probablement renforcée par plusieurs spécificités des apprenants d'une langue étrangère. Il est en fait bien établi dans la littérature que :

- la diversité lexicale des textes d'apprenants s'accroît en fonction de l'amélioration de leur compétence dans la langue en question (Cumming, 2001; McCarthy, 2005);
- la diversité lexicale est un excellent prédicteur de l'évaluation d'un texte par des experts (Yu, 2010);
- par construction, le cosinus interphrases, même obtenu par ASL, est d'autant plus élevé que deux phrases contiennent un plus grand nombre de mots communs.

- 29 Il s'ensuit que plus le vocabulaire d'un texte est diversifié, moins le cosinus est élevé et meilleure est l'évaluation de ce texte par des experts. Dans la présente recherche, cette interprétation est justifiée par l'analyse des corrélations entre une

mesure de diversité lexicale, le *Type/Token Ratio* (TTR) (le rapport entre le nombre de mots différents et le nombre total de mots dans un texte ; plus il est grand et plus le vocabulaire est riche), et les deux autres mesures. En effet, la corrélation entre le TTR et le cosinus moyen est de  $-0,45$  et la corrélation entre ce même TTR et l'évaluation des experts est de  $0,43$ .

30 La longueur moyenne des segments est également très significativement liée à la qualité des essais, mais, dans ce cas, la relation est positive : meilleur est un essai, plus longs sont les segments et, inversement, moins bon est un essai, plus courts sont les segments. Le sens de cette relation correspond à l'hypothèse dérivée des différences de stratégies rédactionnelles liées à l'expertise. Il s'agit toutefois d'un premier résultat dont les limitations sont discutées dans la conclusion. L'annexe 1 essaie de rendre cette relation plus concrète en présentant deux textes qui ont été évalués d'une manière totalement opposée par les experts puisque le premier a été jugé B1- et le second C2. Ils sont découpés en fonction des phrases telles que déterminées par *TreeTagger* et les lignes signalent les ruptures détectées par la procédure de segmentation thématique. On donne aussi le sujet de l'essai, son cosinus interphrases moyen et la longueur moyenne des segments.

31 On remarque tout d'abord de nombreuses différences entre les deux essais tant au niveau de la richesse lexicale que de la complexité syntaxique ou de la correction de la langue, qui expliquent que le second a été jugé comme très nettement supérieur au premier. Comme évoqué ci-dessus, la différence en termes de richesse lexicale explique probablement une partie de la différence entre les cosinus moyens de ces deux textes.

32 En ce qui concerne le second indice étudié, on note que les segments détectés par la procédure automatique semblent relativement bien positionnés, même si certaines ruptures peuvent être avantageusement décalées d'une ou deux phrases. Le premier texte, le plus mal évalué et celui dont les segments sont les plus courts, s'accorde bien avec la stratégie des connaissances rapportées. Le second, en revanche, présente la structure nettement plus diffuse censée résulter de l'application de la stratégie des connaissances transformées. Si l'on parvient à y détecter différents thèmes, le texte glisse de l'un à l'autre sans rupture bien nette et les interconnexions sont nombreuses comme la mention à deux reprises, séparées par plus de neuf phrases, de la mère de l'auteur.

## 7. Discussion et conclusion

33 L'objectif principal de cette recherche est d'évaluer la possibilité d'employer des indices dérivés d'une analyse automatique de la cohésion lexicale pour prédire la qualité de textes. Deux indices ont été considérés : le cosinus moyen entre les phrases contiguës calculé sur la base d'un espace sémantique obtenu par ASL et la longueur moyenne des segments qu'un algorithme détecte dans un texte sur la base de la cohésion lexicale également estimée par une ASL.

- 34 L'analyse du premier indice confirme les observations d'études récentes qui ont mis en évidence une absence de relation ou même une relation négative entre le cosinus moyen interphrases et la qualité d'un texte. *A priori* contre-intuitif, ce résultat semble pouvoir être expliqué par la double relation qu'entretient la diversité lexicale avec la qualité d'un texte et le cosinus interphrases mesuré par l'ASL. Plus un texte repose sur un vocabulaire diversifié, meilleure est l'évaluation des experts et plus faible est le cosinus interphrases. Si l'on peut imaginer différents artifices pour atténuer l'impact sur le cosinus de la répétition de mots dans des phrases contiguës, ces remèdes auraient pour inconvénient majeur de dénaturer les textes et donc la mesure de la cohésion. La présente étude suggère plutôt d'abandonner cet indice parce qu'il présente le défaut majeur d'être intrinsèquement intertextes. Ce qu'on analyse, c'est une estimation de la quantité globale de cohésion dans un texte et non la distribution de la cohésion à l'intérieur d'un texte. Il est important de retenir que ces résultats n'impliquent pas que l'ASL soit inefficace à mesurer la cohésion lexicale puisque cette mesure de cohésion lexicale intervient dans l'algorithme de segmentation thématique, mais plutôt que le cosinus moyen interphrases mesure indirectement la richesse lexicale d'un texte et que c'est celle-ci qui est liée à la qualité.
- 35 Pour dépasser cette limitation, une nouvelle mesure de qualité basée sur la cohésion et dérivée d'un modèle de l'expertise rédactionnelle est proposée. S'appuyant sur la tendance des rédacteurs les moins experts à produire un grand nombre de petits paquets de phrases très cohérentes mais sans lien bien établi entre eux, alors que la structure des textes des rédacteurs experts est beaucoup plus diffuse, elle se base sur la longueur moyenne des segments détectés dans les textes par une procédure automatique. Les analyses mettent en évidence la relation positive postulée : les textes contenant les segments les plus courts sont aussi les textes les moins bien évalués par les experts. On notera que la corrélation pour ce second indice (0,26) est, en valeur absolue, plus faible que celle obtenue pour le premier, le cosinus moyen interphrases (-0,31). Ne devrait-on pas en conclure que le cosinus moyen est plus utile pour prédire la qualité d'un texte ou qu'il faudrait combiner les deux indices ? L'interprétation proposée ci-dessus de la relation entre le cosinus moyen et la qualité d'un texte mène à rejeter ces options : le cosinus moyen prédit la qualité parce qu'il est lié à la richesse lexicale. Il vaut donc mieux employer directement celle-ci ou la combiner avec la longueur des segments si on veut améliorer la prédiction.
- 36 Ces résultats ont été obtenus par l'analyse d'un nombre relativement important de textes d'apprenants d'une langue étrangère (N=223), tous issus du corpus le plus renommé dans ce domaine de recherche, l'*International Corpus of Learner English*. Les textes ont été évalués par deux experts et l'accord entre ceux-ci est relativement élevé ( $r=0,69$ ). Cette valeur est du même ordre de grandeur que celles qui ont été obtenues dans d'autres études alors que celles-ci ont eu recours, contrairement à la présente recherche, à une procédure d'entraînement des juges afin d'uniformiser leurs points de vue (Crossley, Roscoe et McNamara,

2011). Ces deux éléments plaident en faveur de la fiabilité des conclusions de la présente recherche. En revanche, celles-ci sont limitées au cas spécifique des textes produits par des apprenants d'une langue étrangère dont on sait qu'ils font face à une tâche particulièrement coûteuse en ressources cognitives (Galbraith, 2009). Il est donc indispensable de reproduire cette étude sur d'autres corpus, tout particulièrement des textes écrits en langue maternelle par des rédacteurs de différents âges. Il serait aussi intéressant d'analyser d'une manière approfondie les désaccords entre les experts qui ont évalué les textes, puisque la corrélation entre leurs jugements est élevée, mais non parfaite. Plus précisément, on peut se demander si ce ne sont pas les mêmes textes qui posent problème aux experts et à la technique d'évaluation automatique.

37 Ces résultats ont également été obtenus grâce à une technique de segmentation thématique considérée comme une des plus efficaces, mais néanmoins imparfaite (voir Bestgen et Piérard [2006] pour une analyse de son efficacité dans le cas particulièrement complexe d'articles du journal *Le Monde*). Évaluer d'autres algorithmes et surtout combiner plusieurs algorithmes afin de profiter des avantages de chacun permettraient peut-être d'améliorer les performances du segmenteur.

38 Les commentaires qui précèdent soulignent plusieurs limitations de cette recherche, justifiant ainsi des investigations complémentaires. Il serait également utile de s'interroger à propos du choix de l'ASL pour estimer la cohésion lexicale. Cette technique n'est pas, tant s'en faut, la seule permettant d'extraire un espace sémantique sur les bases de l'analyse des cooccurrences dans un corpus de textes (voir par exemple, Lund et Burgess, 1996; Griffiths, Steyvers et Tenenbaum, 2007; Jones et Mewhort, 2007). L'ASL a été choisie parce que c'est la technique qui est la plus utilisée pour mesurer la cohésion entre des phrases et parce que les études qui comparent les espaces sémantiques construits par différentes approches ne mettent habituellement pas en évidence de différences importantes entre elles (Lee, Song et Kim, 2010). Ces études ont toutefois presque toujours comparé les performances dans des tâches lexicales. Étendre ces comparaisons à l'analyse de la cohésion lexicale serait donc intéressant. Plus généralement, cette recherche souligne l'intérêt qu'il y aurait à revisiter les travaux déjà anciens sur l'expertise rédactionnelle dans l'optique de l'évaluation automatique de textes et selon le point de vue des techniques modernes de TAL. La présente étude, par son caractère exploratoire, n'est qu'un premier pas dans cette direction.

## Remerciements

39 Yves Bestgen est chercheur qualifié du F.R.S-FNRS. Les analyses de la cohésion lexicale mesurée par le cosinus interphrases ont été présentées aux journées internationales d'Analyse statistique des données textuelles, Rome, juin 2010 (JADT 2010) (Bestgen, Lories et Thewissen, 2010).

## Références

- BERRY, M. *et al.* 1993. *SVDPACKC (Version 1.0): User's Guide*. Rapport technique CS-93-194. Knoxville: University of Tennessee, Computer Science Department.
- BESTGEN, Y. 2006. Improving Text Segmentation Using Latent Semantic Analysis: A Reanalysis of Choi, Wiemer-Hastings, and Moore (2001). *Computational Linguistics* 32 (1): 5-12.
- BESTGEN, Y. et CABIAUX, A.F. 2002. L'analyse sémantique latente et l'identification des métaphores. In *TALN 2002/RECITAL 2002 (9<sup>e</sup> conférence annuelle sur le Traitement automatique des langues naturelles, Nancy, 24-27 juin 2002)*. Rocquencourt: INRIA: 331-337.
- BESTGEN, Y., DEGAND, L. et SPOOREN, W. 2006. Towards Automatic Determination of the Semantics of Connectives in Large Newspaper Corpora. *Discourse Processes* 41 (2): 175-193.
- BESTGEN, Y., LORIES, G. et THEWISSEN, J. 2010. Using Latent Semantic Analysis to Measure Coherence in Essays by Foreign Language Learners? In S. BOLASCO, I. CHIARI et L. GIULIANO (éd.), *JADT 2010: 10th International Conference on Statistical Analysis of Textual Data*. 385-395. En ligne à l'adresse suivante : [http://www.leonline.it/ledonline/JADT-2010/allegati/JADT-2010-0385-0396\\_139-Bestgen.pdf](http://www.leonline.it/ledonline/JADT-2010/allegati/JADT-2010-0385-0396_139-Bestgen.pdf).
- BESTGEN, Y. et PIÉRARD, S. 2006. Comment évaluer les algorithmes de segmentation automatique? Essai de construction d'un matériel de référence. In C. FAIRON et A. DISTER (éd.), *Verbum ex machina. Actes de la 13<sup>e</sup> conférence sur le Traitement automatique des langues naturelles*. Louvain-la-Neuve: Presses universitaires de Louvain. Vol. I: 407-414.
- BURSTEIN, J. 2009. Opportunities for Natural Language Processing Research in Education. In A. GELBUKH (éd.), *Computational Linguistics and Intelligent Text Processing: 10th International Conference, CICLing 2009, Mexico City, Mexico, March 1-7, 2009: Proceedings*. Berlin: Springer: 6-27.
- BURSTEIN, J., CHODOROW, M. et LEACOCK, C. 2004. Automated Essay Evaluation: The Criterion Online Writing Service. *AI Magazine* 25 (3): 27-36.
- CHANQUOY, L. et ALAMARGOT, D. 2003. Mise en place et développement des traitements rédactionnels: le rôle de la mémoire de travail. *Le langage et l'homme* 38 (2): 171-190.
- CHOI, F. 2000. Advances in Domain Independent Linear Text Segmentation. In J. WIEBE (éd.), *Proceedings of the 1st Meeting of the North American Chapter of the Association for Computational Linguistics*. Stroudsburg: Association for Computational Linguistics: 26-33. En ligne à l'adresse suivante : <http://www.aclweb.org/anthology-new/A/A00/A00-2004.pdf>.
- CHOI, F., WIEMER-HASTINGS, P. et MOORE, J. 2001. Latent Semantic Analysis for Text Segmentation. In D. YAROWSKY (éd.), *Proceedings of the 2001 Conference on Empirical Methods in Natural Language Processing*. Stroudsburg: Association for Computational Linguistics: 109-117. En ligne à l'adresse suivante : <http://www.aclweb.org/anthology-new/W/W01/W01-0514.pdf>.
- COUNCIL OF EUROPE 2001. *Common European Framework of Reference for Languages: Learning, Teaching, Assessment*. Cambridge – New York – Oakleigh: Cambridge University Press.

- CROSSLEY, S.A. et MCNAMARA, D.S. 2012. Predicting Second Language Writing Proficiency: The Roles of Cohesion and Linguistic Sophistication. *Journal of Research in Reading* 35 (2) : 115-135.
- CROSSLEY, S.A., ROSCOE, R. et MCNAMARA, D.S. 2011. Predicting Human Scores of Essay Quality Using Computational Indices of Linguistic and Textual Features. In G. BISWAS *et al.* (éd.), *Artificial intelligence in education : Proceedings. 15th International Conference, AIED 2011, Auckland, New Zealand, June 28-July 2, 2011*. Berlin : Springer : 438-440.
- CUMMING, A. 2001. Learning to Write in a Second Language: Two Decades of Research. *International Journal of English Studies* 1 (2) : 1-23.
- DEANE, P. et QUINLAN, T. 2010. What Automated Analyses of Corpora Can Tell Us about Students' Writing Skills. *Journal of Writing Research* 2 (2) : 151-177.
- FERRET, O. 2002. Using Collocations for Topic Segmentation and Link Detection. In *COLING 2002: The 19th International Conference on Computational Linguistics*. Stroudsburg : Association for Computational Linguistics : 260-266. En ligne à l'adresse suivante : <http://aclweb.org/anthology-new/C/Co2/Co2-1033.pdf>.
- FOLTZ, P.W. 2007. Discourse Coherence and LSA. In T.K. LANDAUER *et al.* (éd.), *Handbook of Latent Semantic Analysis*. Mahwah (N.J.) : L. Erlbaum : 167-184.
- FOLTZ, P.W., KINTSCH, W. et LANDAUER, T.K. 1998. The Measurement of Textual Coherence with Latent Semantic Analysis. *Discourse Processes* 25 (2-3) : 285-307.
- GALBRAITH, D. 2009. Cognitive Models of Writing. *German as a Foreign Language* 2-3 : 7-22.
- GRAESSER, A.C. et MCNAMARA, D.S. 2011. Computational Analyses of Multilevel Discourse Comprehension. *Topics in Cognitive Science* 3 (2) : 371-398.
- GRANGER, S. *et al.* (éd.) 2009. *International Corpus of Learner English. Version 2 (Handbook + CD-ROM)*. Louvain-la-Neuve : Presses universitaires de Louvain.
- GRANGER, S. et TYSON, S. 1996. Connector Usage in the English Essay Writing of Native and Non-Native EFL Speakers of English. *World Englishes* 15 (1) : 17-27.
- GRIFFITHS, T.L., STEYVERS, M. et TENENBAUM, J.B. 2007. Topics in Semantic Representation. *Psychological Review* 114 (2) : 211-244.
- HALLIDAY, M.A.K. et HASAN, R. 1976. *Cohesion in English*. Londres : Longman.
- HEARST, M.A. 1997. TextTiling: Segmenting Text into Multi-Paragraph Subtopic Passages. *Computational Linguistics* 23 (1) : 33-64.
- JONES, M.N. et MEWHORT, D.J.K. 2007. Representing Word Meaning and Order Information in a Composite Holographic Lexicon. *Psychological Review* 114 (1) : 1-37.
- KEHAGIAS, A., PAVLINA, F. et PETRIDIS, V. 2003. Linear Text Segmentation Using a Dynamic Programming Algorithm. In *Proceedings of the 10th Conference of the European Chapter of the Association for Computational Linguistics*. Stroudsburg : Association for Computational Linguistics : 171-178. En ligne à l'adresse suivante : <http://aclweb.org/anthology-new/E/E03/E03-1058.pdf>.
- KOZIMA, H. 1993. Text Segmentation Based on Similarity between Words. In *31st Annual Meeting of the Association for Computational Linguistics. Proceedings of the Conference: 22-26 June 1993, Ohio State University, Columbus, Ohio, USA*. Morristown (N.J.) : Association for Computational Linguistics : 286-288.

- LANDAUER, T.K. *et al.* (éd.) 2007. *Handbook of Latent Semantic Analysis*. Mahwah (N.J.): L. Erlbaum.
- LANDAUER, T.K., FOLTZ, P.W. et LAHAM, D. 1998. An Introduction to Latent Semantic Analysis. *Discourse Processes* 25 (2-3): 259-284.
- LANDAUER, T.K., LAHAM, D. et DERR, M. 2004. From Paragraph to Graph: Latent Semantic Analysis for Information Visualization. *Proceedings of the National Academy of Science* 101 (Suppl. 1): 5214-5219.
- LANDAUER, T.K. et PSOTKA, J. 2000. Simulating Text Understanding for Educational Applications with Latent Semantic Analysis: Introduction to LSA. *Interactive Learning Environments* 8 (2): 73-86.
- LAPATA, M. et BARZILAY, R. 2005. Automatic Evaluation of Text Coherence: Models and Representations. In A. SAFFIOTTI et L.P. KAELBLING (éd.), *IJCAI-05: Proceedings of the 19th International Joint Conference on Artificial Intelligence, Edinburgh, Scotland, July 30-August 5, 2005*. Denver: Professional Book Center: 1085-1090.
- LEE, S., SONG, J. et KIM, Y. 2010. An Empirical Comparison of Four Text Mining Methods. *Journal of Computer Information Systems* 51 (1): 1-10.
- LUND, K. et BURGESS, C. 1996. Producing High-Dimensional Semantic Spaces from Lexical Co-Occurrence. *Behavior Research Methods, Instruments & Computers* 28 (2): 203-208.
- MAGLIANO, J.P. et GRAESSER, A.C. 2012. Computer-Based Assessment of Student-Constructed Responses. *Behavior Research Methods* 44 (3): 608-621.
- MANNING, C. et SCHÜTZE, H. 1999. *Foundations of Statistical Natural Language Processing*. Cambridge – Londres: MIT Press.
- MCCARTHY, P.M. 2005. *An Assessment of the Range and Usefulness of Lexical Diversity Measures and the Potential of the Measure of Textual, Lexical Diversity*. Thèse de doctorat. University of Memphis.
- MCCUTCHEN, D. 1996. A Capacity Theory of Writing: Working Memory in Composition. *Educational Psychology Review* 8 (3): 299-325.
- MILLER, T. 2003. Essay Assessment with Latent Semantic Analysis. *Journal of Educational Computing Research* 29 (4): 495-512.
- MILTSAKAKI, E. et KUKICH, K. 2004. Evaluation of Text Coherence for Electronic Essay Scoring Systems. *Natural Language Engineering* 10 (1): 25-55.
- MORRIS, J. et HIRST, G. 1991. Lexical Cohesion Computed by Thesaural Relations as an Indicator of the Structure of Text. *Computational Linguistics* 17 (1): 21-48.
- PAGE, E.B. 1966. The Imminence of Grading Essays by Computer. *Phi Delta Kappan* 48 (5): 238-243.
- PIÉRARD, S. et BESTGEN, Y. 2008. Use of Temporal Adverbials as Segmentation Discourse Markers by Second Language Learners. *Archives de Psychologie* 73: 209-230.
- REHDER, B. *et al.* 1998. Using Latent Semantic Analysis to Assess Knowledge: Some Technical Considerations. *Discourse Processes* 25 (2-3): 337-354.
- SCARDAMALIA, M. et BEREITER, C. 1987. Knowledge Telling and Knowledge Transforming in Written Composition. In S. ROSENBERG (éd.), *Advances in Applied Psycholinguistic. Volume 2: Reading, Writing and Language Learning*. Cambridge: Cambridge University Press: 142-175.

- SCHMID, H. 2003. Probabilistic Part-of-Speech Tagging Using Decision Trees. In H.L. SOMERS et D.B. JONES (éd.), *New Methods in Language Processing*. Londres: Routledge: 154-164.
- THEWISSEN, J. 2012. *Accuracy across Proficiency Levels and L1 Backgrounds: Insights from an Error-Tagged EFL Learner Corpus*. Thèse de doctorat. Université catholique de Louvain, Centre for English Corpus Linguistics.
- YEUNG, L. 2009. Use and Misuse of «Besides»: A Corpus Study Comparing Native Speakers' and Learners' English. *System* 37 (2): 330-342.
- YU, G. 2010. Lexical Diversity in Writing and Speaking Task Performances. *Applied Linguistics* 31 (2): 236-259.

## Annexe 1. Deux exemples de textes issus du corpus analysé

### Texte 1

Sujet de l'essai: Crime and conviction.

Évaluation CECR: B1-; cosinus moyen: 0,30; longueur moyenne d'un segment: 4,4 phrases; TTR: 0,35.

A crime is a very serious violation of the law, this definition is very general but it is as broad, because there are a lot of kinds of crimes.

Some examples of crimes could be these ones:

a) Murder: in this kind of crime somebody is killed and perhaps, this is the worst crime because if something is robbed it can be recovered but, if someone is killed, that person is not going to live again.

The most common reason to kill a person is by revenge, but it is true that today a lot of people is killed for others causes as money, drugs...

b) Robe: in large cities everyday a lot of people is robbed though in little towns and villages it can happens the same.

---

Nowadays, people have serious problems with their jobs or they are unemployed and haven't got a lot of money to live so they must rob to continue living.

This is an important reason why people rob but, the main problem is that young men have drugs and it is very expensive.

If they want to continue having drugs, they must rob because they haven't got enough money to buy it.

The majority of thieves are people who haven't got a stable job and that have drugs so the problem is frightful for the thief and for normal citizens who are robbed.

---

c) Rape: in these last years, it can be watched in television that many young girls are raped not only in the street but even in their own houses street door.

This is very common so, young girls are worry when they are alone in the street by night.

---

The rapists must be «mad» people, because a man which hasn't got problems in his mind doesn't rape anybody.

To pay the crimes, criminals are put in prison during a time that is different according to the crime that has been done.

This is the only solution that government has studied to pay the crime but it is not a good manner of resolve a problem.

When many criminals that have been in prison serve their sentences they are left into society as they were when they committed the crime so, it is sure that many of the criminals are going to kill and robe again.

In this way we can not say that the problem is solved.

The correct manner of aliminate the problems is that criminals have psychological therapies with doctors because, the problem is not in their hands but in their minds.

---

According to what has been explained, the crime doesn't pay in our society but, it does not mean that in other time or place it happens the same.

The meaning of these words is that christian people think that if they have done something bad in their present life, in their future life arter they die, they are going to pay the crimes that they committed in their past life.

This is a very important idea because in Spain, for instance, the majority of people are christians but, there are a lot of criminals so, there is a very great contradiction.

If this is so, it is because people have the necessity of robe and, if they can not do other thing to live, they forget the religion and robe and kill.

This is very painful but it is sure that crime does pay in other life.

## Texte 2

Sujet de l'essai: Clothes maketh man?

Évaluation CECR: C2; cosinus moyen: 0,12; longueur moyenne d'un segment: 20 phrases; TTR: 0,45.

I'm not quite sure if I do agree to this saying.

Well, clothes are often responsible for the first impression one gets of a person, responsible for many prejudices.

Isn't it a bit dangerous to infer the character of a person from his immaculately ironed white shirt or his loose-fitting lilac corduroy breeches?

There is, for exemple, the skinny tall young man walking to the lecture-hall with his brown brief-case.

He's wearing a neat white shirt, a grey suit and glossy black patent leather shoes.

Well, he's studying science of industrial management.

But you've known it anyway, haven't you?

And there is this young man with his long fair hair bound together to a plait, wearing loosefitting lilac trousers, a big knitted colourful jumper, home-made socks and Birkenstock-shoes.

He's sitting on a bench in the public park, observing the children playing in the sand-pit.

Well, he should rather sit behind an information stand of Greenpeace and roll a cigarette, shouldn't he?

Again in the park there is a jogger, aged about 25, running around in his neon yellow and orange track-suit and his new white sports shoes.

Moreover, he's got a red, orange and yellow striped headband, wristlets in the same colours and a walkman hanging round his neck.

How arrogant he must be!

But wait a minute, this man reminds me of Thomas.

Yes, I also met him in a park, and he was such a nice guy.

And I'm sure there's also a nice student of industrial management and a nice Birkenstock-wearer.

Even the lady in the café who is sitting at the table next to yours, wearing a tight chequered black and white mini-skirt and a coat of the same cloth, a grey silk blouse with an awful lot of ruches, grey silk stockings, high-heeled shoes and an elegant hat – and even the tourist in Australia, wearing a pistachio-coloured shirt, khaki shorts and a white hat in a safari-look, not to forget the camera hanging round his neck – yes, and even the aged tramp sitting under the bridge in his old and dirty black cotton trousers, black boots and his thick grey coat, which hopefully is warm – they all, believe it or not, they all might be very nice people.

Maybe I'm a burnt child in this context.

Since the time when I started to decide myself which clothes to wear I've been in trouble with my mother.

No matter if I was wearing clinging black jeans and my grandfather's chequered shirt or my ruby-coloured long skirt, a grey knitted jumper, grey tights and black boots.

As she always wanted a neat girl, she had to carp at my clothes on principle.

And I never could understand how something like clothes can be so important for the relationship between two people.

I always felt guilty in a way, when I wore something she didn't like.

But last year I went to Ireland for a few months to live in a Camphill Community.

What I enjoyed the most from the beginning to the end was that everybody could wear whatever he wanted.

There was Gail Richards, for exemple, who was about 25 years old.

She wasn't very tall, indeed, but she was wearing huge trousers, shirts and jumpers all the time.

And I had never heard anybody saying: «Look at Gail!» or «Gail, you really can't dress like this.

What will the people think of you?»

It was Gail as well, who showed me the second-hand-box in the attic.

Among colourful waistcoats, torn and washed out jeans, old-fashioned blouses and home-made Aran jumpers I discovered a loose fitting glaring orange pair of leggings and a big white sweatshirt with something American printed on it: «King's Canyon – Sequoia Park – Fifth Winter Trophy» – in bright colours, of course.

I found these clothes wonderful and always put them on before I went milking the cows.

If my mother had seen me, she would have had a heart attack or something like that.

I've learned a lot in Ireland.

Especially, I've learned to dress in the way I liked without thinking of other people's opinions.

But still, there are some situations where I conform to the social rules concerning clothes.

One of them is the Staatsexamen.

I think I'll do it the same way as my fellow-lodger did two years ago: At home, while she was learning, she was wearing old grey gym trousers and a worn out jumper which she permanently had pulled over her knees while sitting at the desk.

But for the exams she put on her new natural-coloured linen-trousers, a smart cotton pullover in the same tone and a mossy green silk scarf.

Clothes maketh man, you know.