

Using latent semantic analysis to measure coherence in essays by foreign language learners?

Yves Bestgen ¹, Guy Lories ², Jennifer Thewissen ¹

¹ UCL/CECL –Place Blaise Pascal, 1 – B-1348 Louvain-la-Neuve –Belgique

² UCL/PSP –Place du cardinal Mercier, 10 – B-1348 Louvain-la-Neuve –Belgique

Abstract

This paper compares human and LSA-based coherence ratings of 223 essays written by learners of English as a foreign language at various levels of linguistic proficiency. While the LSA-based ratings correlate positively – as expected – with a measure of lexical variety, a negative correlation was found between these ratings and the human grading. This is interpreted as resulting from specific writing strategies used by authors of different proficiency levels but also from the fact that human graders were sensitive to a wider range of linguistic cohesive devices.

Keywords: Coherence and cohesion, Latent Semantic Analysis, Foreign Language Learners

1. Introduction

For a few years now, Latent Semantic Analysis (LSA) has been at the heart of an increasing number of studies in psychology and educational sciences (Landauer et al., 2007). This technique rests on the singular value decomposition theorem which is central to the factorial methods of data analysis (Lebart et al., 2000). LSA is used to build a very large “semantic space” from a text corpus through a statistical analysis of word co-occurrences. This semantic space is then used to estimate the similarity of meaning between sentences, paragraphs or even texts. Among the many applications of LSA, automatic essay grading is certainly one of the most important for psychology and education. In this context LSA has been used to automatically compare students’ essays to a “golden standard” (Landauer and Psootka, 2000; Rehder et al., 1998; van Bruggen et al., 2004). It has also been proposed to use LSA to evaluate more specific elements like coherence or conceptual cohesion in essays (Bestgen et al., 2006; Folz, 2007; Foltz et al., 1998; Wiemer-Hastings and Graesser, 2000). One way of doing this is to measure the thematic similarity of adjacent segments of a text. Another way is to compute the similarity between each segment and the text as a whole. Such indices can be used to rank the essays according to their quality but can also help to identify the segments that are less coherent, hence guiding the revision process. This kind of analysis is made possible thanks to tools as for example Coh-Metrix, a Web-based software tool which analyzes dozens of cohesion measures such as coreferential cohesion and causal cohesion (Graesser et al., 2004).

Up to now these techniques have often been used on texts whose coherence had been manipulated to allow for the testing of various hypotheses about the comprehension process (Foltz et al., 1998; McNamara et al., 2006; 2007) or to compare the level of coherence between or within textbooks (Dufty et al., 2006). These techniques have almost never been used to assess

discourse produced in relatively natural situations although there are some notable applications like measuring a shift in topic during conversation between team members (Folz, 2007) or the detection of abnormal coherence patterns in the discourse of schizophrenic patients (Elvevåg et al., 2006).

Very recently, Crossley et al. (2008a; 2008b) applied LSA as implemented in Coh-Metrix to a longitudinal study of coherence in the discourse of learners of English as a Foreign Language (EFL). The authors observed that coherence between adjacent utterances spoken by an L2 learner improved throughout an intensive English program. This evolution occurs in parallel with an increase in lexical proficiency as measured by an index of lexical diversity. These are encouraging results, not only because they provide further validation of the technique but also because they help assess the coherence and cohesion in the production of non-native speakers. A series of studies has shown that, compared to native speakers, non-native speakers have difficulties in mastering cohesion (Hinkel, 2001; Silva, 1993). However, these studies are mainly based on the analysis of explicit cohesion devices such as connectives and pronouns. In addition, the number of texts analyzed is often very small because analyzing a text for cohesion is extremely time-consuming (Hinkel, 2005). The use of automatic techniques like LSA makes up for this double caveat, allowing for the construction of new measures and opening practical perspectives by making most of the work automatic.

The results presented in Crossley et al. (2008a; 2008b) have some limitations due to sample size (6 learners) and to the learners' low level of English proficiency at the beginning of the study. Moreover, these studies are based on conversations with assessors during elicitation sessions. This probably limits the relevance of the data regarding essay grading and possible generalizations. More critically, quite a few studies on the cohesion of foreign language learner texts do not actually provide direct and unanimous support to the belief that better essays are invariably characterized by a higher level of coherence and cohesion (Castro, 2004; Johnson, 1992; Mojica, 2006). In the case of oral narratives, Strömquist and Day (1993) observed that children improve their use of cohesive devices in their L1 between the ages of 3 and 5, and between 5 and 8, whereas adult L2 learners manage these devices more efficiently at the beginning of their learning than 7 to 9 months later. The drop in performance for adult L2 learners apparently results from their attempt at using more complex lexico-syntactic patterns to the detriment of efficient use of cohesive devices (Strömquist and Day, 1993). Regarding the LSA measure of coherence specifically, Foltz (2007: 177) stresses that in his experience «better essays often tend to have low coherence, because a student must cover multiple facets or examples in a short span of text». Although specific data are not available, it cannot be taken for granted that Crossley et al.'s (2008a; 2008b) results would be replicated in grading essays produced by EFL learners at various levels of proficiency.

This limited amount of data available contrasts with the position taken by the editors of the *Handbook of Latent Semantic Analysis* when, in the foreword, they state that LSA «accurately estimates passage coherence (Landauer et al., 2007: p. X)». We therefore set out to investigate the LSA coherence measures obtained from 223 EFL essays and compared them to the grades given by human experts. We also analyzed two measures of lexical diversity. We had two reasons for doing so: (1) longitudinal analyses have shown that there is a strong increase in lexical diversity in non-native speakers as proficiency improves, and (2) non-native speakers are known for meeting difficulties in this area (McCarthy, 2005; Cumming, 2001). These factors could affect LSA measures of coherence as the more repeated words a text contains, the higher the cosines between sentences tend to be.

2. Method

2.1. Corpus of essays

The essays analyzed in this study are part of the International Corpus of Learner English (ICLE (v1.1) (Granger et al., 2002), a corpus of essays written by intermediate to advanced learners of English as a foreign language. ICLE is the result of over ten years of active collaborative research between a large number of international universities. It contains over two million words from 3.640 learners from 11 different mother tongue backgrounds. Two hundred and twenty three argumentative essays were extracted from three ICLE subcorpora, i.e. 74 essays were taken from the French (FR) component, 71 from the German (GE) component and 78 from the Spanish (SP) component of the learner corpus. The essays had to meet two criteria: each text had to be between 500 and 900 words long and it had to be argumentative in nature. They have been selected within the framework of a PhD project carried out at the Centre for English Corpus Linguistics, Belgium (Thewissen, 2008; forthcoming).

Each learner in ICLE has a learner profile (age, sex, time spent in an English-speaking country, years at university, etc.). A short univariate summary of some of these descriptive variables is presented in Tab. 1. Due to missing data, information is only provided for 212 learners.

<i>Variables</i>	<i>Mean</i>	<i>Std</i>	<i>Min</i>	<i>Max</i>
Age	22.24	2.40	19	31
Years of English at University	3.37	1.20	1	6
Months in English speaking country	3.26	5.33	0	36

Table 1: Descriptive variables from the ICLE learner profile

2.2. Ratings by human experts

Two professionally-trained raters were called upon to grade each of the 223 EFL texts. The raters have both worked at the University of Cambridge Local Examinations Syndicate (UCLES) and have been involved in the assessment of writing for the higher level Cambridge exams. They were, among others, asked to assign a grade for cohesion and coherence based on the following descriptor scale (Tab. 2) devised on the basis of the Common European Framework (Council of Europe, 2001).

<i>Grade</i>	<i>Descriptors</i>
B1	Can link a series of shorter, discrete simple elements into a connected, linear sequence of points
B2	Can use a limited number of cohesive devices to link his/ her utterances into clear, coherent discourse, though there may be some 'jumpiness' in a long contribution.
B2+	Can use a variety of linking words efficiently to mark clearly the relationship between ideas.
C1	Can produce clear, smoothly flowing, well-structured texts, showing controlled use of organizational patterns, connectors and cohesive devices.
C2	Can create coherent and cohesive text making full and appropriate use of a variety of organizational patterns and a wide range of cohesive devices.

Table 2: CEF Descriptors for the cohesion and coherence Sc

It must be noted that the CEF descriptors ask raters to take into account cohesive devices of all types and not only the kind of semantic similarity that is at the heart of coherence or cohesion estimates in LSA (Foltz, 2007; Crossley et al., 2008a).

Raters were asked to judge coherence and cohesion as being either B1, B2, C1 or C2, i.e. the intermediate (B) and advanced (C) levels of proficiency in the CEF. They were allowed to use + or – signs to further distinguish between sublevels.

These categories were recoded as a numerical scale from 0.667 (for B1-) to 4 (for C2; there were no C2+ scores). The correlation between the grades given by two raters on this 11-point scale was 0.69. Overall, both raters completely agreed on a total of 107 texts (48%); they reached near agreement (a maximal difference of one band score e.g. B2-C1 or C1-C2) on 87 texts (39%) and disagreed by more than one band score on 29 texts (13%). Based on these evaluations, a global score was computed using the mean of the two raters. Tab. 3 provides descriptive statistics for the two raters and the mean ratings.

<i>Variables</i>	<i>Mean</i>	<i>Std</i>	<i>Min</i>	<i>Max</i>
Rater 1	2.71	1.05	0.67	4.00
Rater 2	2.33	1.15	0.67	4.00
Mean Ratings	2.52	1.01	0.67	4.00

Table 3: Descriptive statistics for the human ratings

2.3. Automatic essay analyses: LSA and Coh-Metrix

To gather coherence measures between adjacent segments of the essays, we used Latent Semantic Analysis. In a nutshell, LSA rests on the hypothesis that analyzing the contexts in which words occur enables an estimation of their similarity in meaning (Deerwester et al., 1990; Landauer and Dumais, 1997). The first step in the analysis is to construct a lexical table containing an information-theoretic weighting of the frequency of occurrence of a word in each document (i.e. sentence, paragraph or text) included in the corpus. This frequency table is submitted to a Singular Value Decomposition that extracts the most important orthogonal dimensions and, consequently, discards the small sources of variability in term usage. Second, each word is attributed a vector of weights indicating its strength of association with each of the dimensions. This makes it possible to measure the semantic proximity between any two words by using, for instance, the cosine of the angle between the corresponding vectors. Proximity between any two sentences (or any other two textual units), even if these sentences were not present in the original corpus, can be estimated by first computing a vector for each of these units – which corresponds to the weighted sum of the vectors of the words included in it. Next, the cosine between these vectors is computed.

To estimate the proximity between two adjacent segments of each essay, an LSA space was extracted from the *General Reading up to 1st year college* TASA corpus that T.K. Landauer (Institute of Cognitive Science, University of Colorado, Boulder) kindly gave us access to. The version we used contains 44.486 documents, each of which is approximately 290 words long. To extract the semantic space from this corpus, a series of decisions had to be taken. First, the corpus, which had not been lemmatized, was cut into documents. Deleting words that include non standard characters as well as a list of 329 stopwords (e.g. *and*, *be*, *the*, *that*) brought the total number of words down to approximately 6 million words. After lowercasing all the characters, these pre-processing steps resulted in a matrix of 65.194 different terms in 44.486 documents. To build the semantic space proper, the singular value decomposition was computed with the SVDPACKC program (Berry, 1992; Berry et al., 1993). The first 300 singular vectors were retained (Landauer et al., 1998).

Before applying LSA measures, the 223 essays were edited to remove non standard characters and were divided into one sentence segments. Cosines between adjacent sentences were then

computed and averaged over sentences for each essay. Because the size of the text segments used to measure coherence is a critical issue, we followed the suggestion in (Foltz, 2007) to use a moving window procedure that compares one group of sentences to the next group of sentence repeatedly across the text. A moving window of size 3, for instance, implies computing the cosine between the vector based on the first three sentences of an essay and the vector for the next three sentences and then moving both segments one sentence further to compute the next cosine. It goes without saying that the classical “sentence to adjacent sentence” mean cosine is a special case of this procedure with a window size of 1. So in our analyses we used a moving window size of 1 to 5.

We also computed the *Measure of Textual and Lexical Diversity* (MTLD) (McCarthy, 2005) that was used by Crossley et al. (2008b). Broadly speaking, MTLD corresponds to the mean sentence length of textual segments whose type token ratio (TTR: the ratio between the number of unique words (the types) divided by the number of tokens of these words) is 0.71. The higher the number of words, the higher the lexical diversity of a particular text. The main advantage of this measure, compared to a more classical TTR is its independence from text length (McCarthy, 2005).

The same texts were also analyzed using Coh-Metrix (version 2.1), a web-based application developed at the Institute for Intelligent Systems (University of Memphis) (Graesser et al., 2004; Shapiro and McNamara, 2000; McNamara et al., 2008). Two indices were retained from the results produced by Coh-Metrix: LSAASSA and TYPTOKC. The first, which Co-Metrix calls LSAASSA, is used to validate our own analysis. It was computed using the TASA *College Level* semantic space. The second, TYPTOKC, is a type token ratio that has the advantage of relying on content words only.

3. Results

3.1. LSA Measures and Grading

The most important analysis targets the relationship between the ratings given by the human judges and the coherence measures yielded by LSA. Tab. 4 presents the correlations between the raters and the LSA measures as well as a short univariate description of the variables. The mean cosine between two adjacent sentences as calculated by Coh-Metrix (LSAASSA), the same cosine calculated on our version of TASA (LSA1), as well as the cosine based on longer moving windows (LSA2-LSA5) all yield negative correlations which are all statistically significant with a p value smaller than 0.001.

<i>Variables</i>	<i>Mean</i>	<i>Std</i>	<i>Correlation with Grading</i>
LSAASSA	0.17	0.06	-0.36
LSA1	0.18	0.07	-0.31
LSA2	0.27	0.09	-0.28
LSA3	0.33	0.11	-0.27
LSA4	0.37	0.12	-0.26
LSA5	0.41	0.13	-0.25

Table 4: LSA measures: Descriptive statistics and correlation with grading

As expected in view of the increasing length of the text segments (Folz, 2007; Hu et al., 2007), the mean cosines increase along with the size of the moving window. At the same time, the correlation with the grading increases slightly along with the size of the window.

These negative correlations require further analyses. Since similar correlations are obtained with the different LSA measures, we shall only report on the results for LSAASSA (Coh-Metrix sentence to adjacent sentence mean cosines).

The relationship between the human ratings and the LSA cosines is illustrated in Fig. 1. The first sentences of the two essays indicated by arrows in the graph are presented below.

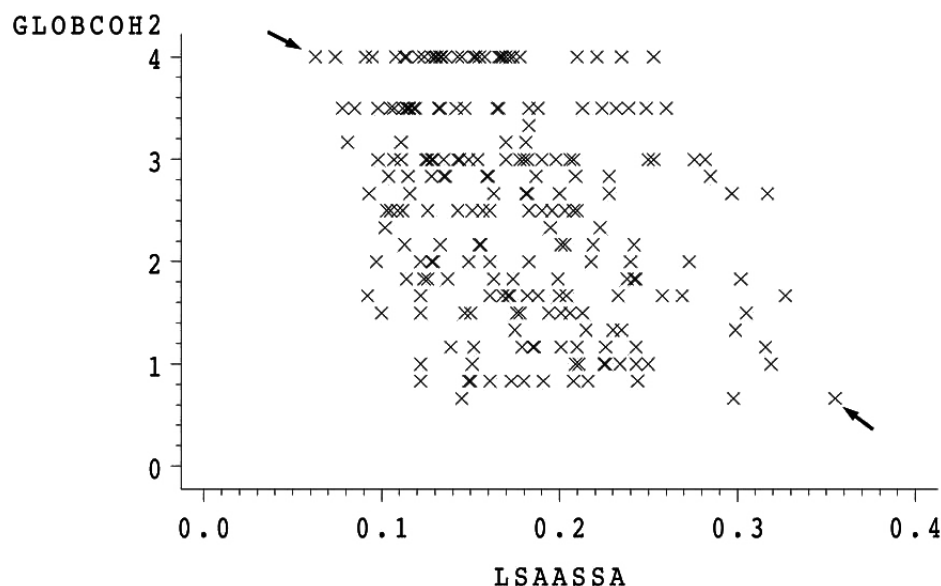


Figure 1: Relation between human ratings (y) and LSA score (x)

CEF Grading : B1- (0.67); LSAASSA : 0.36

Title: Crime does not pay

Nowadays, many crimes are committed which are not punished correctly (I am referring to Spain but I suppose that the situation is very similar in other countries); for this reason, it is said that crime does not pay. This seems true since the only ones that pay are citizens which are the victims. Citizens are fed up with this situation of insecurity and injustice so they demand a better protection on the streets, just trials and right penalties. Although it is difficult to put an end to crime, if the Legal System and the police method was changed, this situation would improve. First of all, penalties are not imposed correctly on several occasions and, moreover, there are not proportion between the crime and the penalty which is imposed by the judge. Whereas some of the biggest crimes; such as, important robberies, kidnappings, kills and rapes are punished with penalties which are too short, some sentences of minor crimes like a robbery to survive are too long...

CEF Grading : C2 (4.00); LSAASSA : 0.06

Title: The single most absurd characteristic of the British is their contradictory attitude towards nudity and the nature of their body

And they realised that they were naked ... and got the shock of their lives. The single most absurd characteristic of the British is their contradictory attitude towards nudity and the nature of their body. On the one hand it seems to be quite fashionable among young students to talk about contraceptives and making love, on the other hand they are extremely bashful. Psychologist Allan Seethrough lately said in a BBC interview: "<>" That is exactly what I experienced as a German student in my year abroad in the UK. It seemed arbitrary to me what kind of things were regarded as indecent and about what kind of things it was the most natural things to talk about. The first day of term all overseas students at the university of Aberystwyth are welcomed in a talk by the dean in the Great Hall. They get a pack of leaflets with all the information you need about scholarships, bursaries, the NHS, arranging*

your courses, etc. and you get a pair of condoms and instructions on how to use them as a welcoming present. I must admit, I was a bit dazzled. Do British people face reality more openly than we Germans?

One only needs to read these two essays to understand the difference in the mean cosines detected by LSA. The first author seems to follow the “knowledge telling strategy”, which has been found to be used by novice writers (Scardamalia and Bereiter, 1987), i.e. basing oneself on the title of the essay to find relevant information and to immediately build sentences. The repeated use of this strategy enables one to write coherent texts. The first excerpt perfectly fits Hinkel’s observation according to which non-native writers attempt «to construct a unified idea flow within the constraints of a limited syntactic and lexical range of accessible linguistic means». (Hinkel, 2001: 128) The second writer, on the other hand, seems to apply the more expert knowledge-transforming strategy (Scardamalia and Bereiter, 1987), which is a problem-solving activity involving extensive planning in order to meet the dual demands of rhetoric and content-related goals. This essay is near-native in quality.

3.2. Lexical Diversity, Grading and LSA Mean Cosine

The indices for lexical diversity calculated for the 223 essays confirm the above observations. As shown in Tab. 5, both indices strongly correlate with the grading, as was the case in other studies (Crossley et al., 2008b; Cumming, 2001). A strong negative correlation appears between the lexical diversity measures and the LSA mean cosine: the less diversified the vocabulary of a text, the higher the LSA mean cosine. This correlation may partly be due to the fact that the number of repeated words increases the cosines between sentences.

As for the cosine measure, the two excerpts presented above are far from each other on this dimension: the B1 excerpt has a TYPTOKC of 0.55 while the C2 excerpt has a 0.78 TYPTOKC

Variables	Mean	Std	Correlation with	
			Grading	LSAASSA
TYPTOKC	0.69	0.08	0.49	-0.49
MTLD	98.85	24.68	0.46	-0.43

Table 5: Lexical diversity: descriptive statistics and correlation with grading and LSA mean cosine

The negative relationship between essay quality and LSAASSA actually contradicts previous results (Crossley et al., 2008a; 2008b) while the two lexical diversity measures seem to behave similarly in both studies. This suggests that the LSA score has a specific meaning in the present context. Actually, if one were to partial out the effect of TYPTOKC from the LSAASSA to essay quality relationship, a small but significant negative correlation between LSAASSA and quality ($r=-0.15$) would remain.

In order to identify the causes of this negative correlation, we tried to consider the behaviour of EFL writers at various levels of proficiency. Fig. 2 shows the relationship between the TYPTOKC and LSAASSA for the B1 and C2 essays. C2 essays have both a low LSAASSA score and a higher TYPTOKC score, suggesting that they make use of their higher lexical variety to shift from topic to topic while using a specific vocabulary for each. B1 essays show the inverse pattern, i.e. smoother transitions and less variety. The B1 authors may simply not have the lexical means to clearly separate the thematic segments. The data for the B2 and C1 groups fall somewhere between the B1 and C2 groups and were left out for legibility.

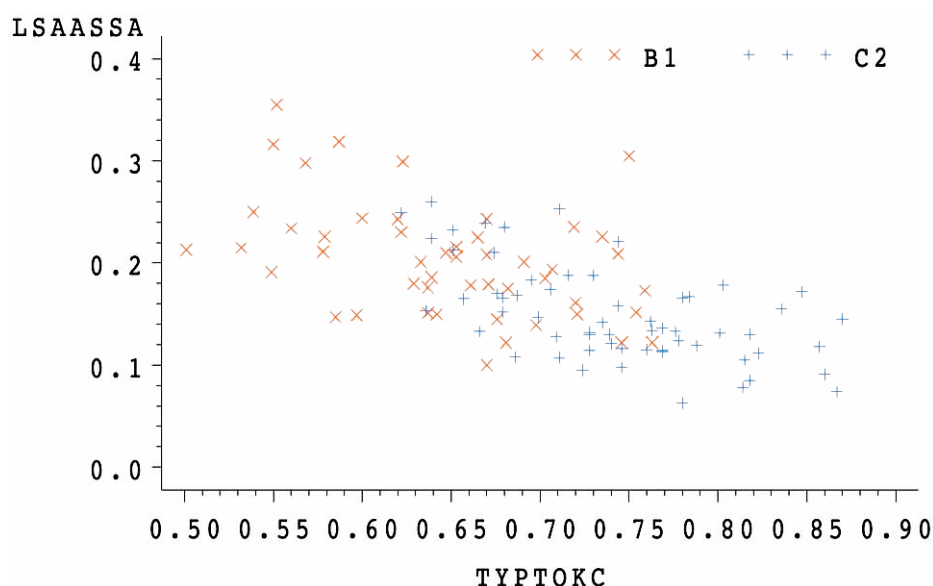


Figure 2: Relation between lexical variety and LSA score

3.3. Grading and Learner Profile

The graph is in agreement with the significant, but not very strong correlation observed between the human rating and the mean cosine calculated by LSA. This result is not surprising given that other factors than the one measured by LSA influence the cohesion and coherence of a text and that many other factors than cohesion and coherence influence the grading of a text, especially texts written by EFL learners. To get a better idea of the genuine or anecdotic nature of these correlations, we compared them to the correlation between the grading and two variables taken from the learner profile which may partly influence the learners' proficiency in English: the number of years of English at University and the number of months spent in an English speaking country. They both correlate at 0.33 with the grading. This value is comparable to the value for the LSA mean cosine and is lower than the result obtained for lexical diversity.

4. Conclusion

The main aim of the present research was to determine whether LSA measures of textual coherence are related to the grades for cohesion and coherence given to ELF essays by expert raters. Our analyses have yielded significant but negative correlations: the higher the LSA coherence indices, the lower the human rating. This tallies with the observation by Foltz (2007), according to which it is not necessarily the best essays which have the higher intersentence cosines. Our results nevertheless contradict those by Crossley et al. (2008a; 2008b). The exact reason for the discrepancy between the two studies is hard to pinpoint. Both studies have indeed used very different methodologies and further research would be necessary here. There are three main discrepancies between the two studies: first, learners in Crossley et al. (2008a; 2008b) were at a very low level of proficiency in English at the beginning of the study, while our sample includes intermediate to advanced learners. Second, analyses in Crossley et al. (2008a; 2008b) target LSA coherence between adjacent utterances in L2 speech while we studied the coherence of adjacent sentences in argumentative written essays. Finally and perhaps more importantly, we need to consider the differences in assessment procedures. While the longitudinal study in Crossley et al. (2008a; 2008b) relied on a proficiency increase due to an intensive learning programme, our analyses involve the judgements of expert human raters.

The crucial point to bear in mind is that the human judges and LSA may tap a different form of coherence or cohesion. In addition, as the CEF uses these two terms rather loosely and lays particular emphasis on a range of cohesive devices, the raters may well have paid more attention to linguistically-marked textual cohesion (*e.g.*, connectives) than to semantic coherence.

The analysis of the relationship between the lexical diversity in the essays and the grading yields results which are more in keeping with those by Crossley et al. (2008b). The texts which get the lowest grades are also those which display the lowest degree of lexical diversity. The correlation of almost -0.50 between the lexical diversity measure and the LSA cosine calls for further analyses. Part of this correlation is probably due to the fact that a less (lexically) diversified text will tend to have a higher mean cosine. The correlation is also probably the consequence of a major difference between better- and lower-quality essays, *i.e.* the use of the knowledge telling strategy by less able writers. The knowledge telling strategy increases the coherence of a text thanks to a limited use of vocabulary. The next point in our research agendas is to develop finer indices that will make it possible to distinguish the knowledge telling and knowledge transforming strategies.

It must be stressed that the aim of this study was not to develop an automatic grading tool for the EFL essays (Lonsdale and Strong-Krause, 2003). We have focussed exclusively on ratings of cohesion and coherence. Our study shows that LSA results must be interpreted with caution. It is nevertheless crucial to remember that this study does not question the usefulness of LSA as a tool to investigate coherence, but suggests that in specific contexts LSA coherence may actually be consistently higher for lower-quality essays than for higher-quality essays. It also points to the need for detailed studies of coherence in native-speaker texts.

Acknowledgment

Yves Bestgen is Research Associate of the Belgian Fund for Scientific Research (F.R.S-FNRS). We gratefully acknowledge the help of T.K. Landauer with the TASA corpus and the team developing Coh-Metrix for making their tool freely available to the scientific community. We would also like to thank Sylviane Granger for very helpful comments on an earlier version of this paper.

References

- Berry M.W. (1992). Large scale singular value computation. *International Journal of Supercomputer Application*, 6: 13-49.
- Berry M.W., Do T., O'Brien G., Krishna V. and Varadhan S. (1993). SVDPACKC: Version 1.0 user's guide. Tech. Rep. CS-93-194, University of Tennessee, Knoxville, TN, October 1993.
- Bestgen Y., Degand L. and Spooren W. (2006). Towards automatic determination of the semantics of connectives in large newspaper corpora. *Discourse Processes*, 41: 175-193.
- Castro C. (2004). Cohesion and the social construction of meaning in the essays of Filipino college students writing in L2 English. *Asia Pacific Education Review*, 5: 215-225.
- Council of Europe (2001). *Common European Framework of Reference for Languages: Learning, Teaching, Assessment*. Cambridge, UK: Cambridge University Press.
- Crossley S.A., Salsbury T., McCarthy P.M. and McNamara D.S. (2008a). LSA as a measure of coherence in second language natural discourse. In *Proceedings of the 30th annual conference of the Cognitive Science Society*, pp.1906-1911.

- Crossley S.A., Salsbury T., McCarthy P.M. and McNamara D.S. (2008b). Using latent semantic analysis to explore second language lexical development. In *Proceedings of the 21st International Florida Artificial Intelligence Research Society Conference*, pp. 136-141.
- Cumming A. (2001). Learning to write in a second language: Two decades of research. *International Journal of English Studies*, 1: 1-23.
- Deerwester S., Dumais S.T., Furnas G.W., Landauer T.K. and Harshman R. (1990). Indexing by latent semantic analysis. *Journal of the American Society for Information Science*, 41: 391-407.
- Dufty D.F., Graesser A.C., Louwerse M. and McNamara D.S. (2006). Assigning grade levels to textbook: Is it just readability? In *Proceedings of the 28th Annual Conference of the Cognitive Science Society*, pp. 1251-1256.
- Elvevåg B., Foltz P., Weinberger D. and Goldberg, T. (2006). Quantifying incoherence in speech: An automated methodology and novel applications to schizophrenia. *Schizophrenia Research*, 93: 304-316.
- Foltz P.W. (2007). Discourse coherence and LSA. In Landauer, T., McNamara, D., Simon, D. and Kintsch, W., editors, *Handbook of Latent Semantic Analysis*, Mahwah (NJ): Erlbaum, pp.167-184.
- Foltz P.W., Kintsch W. and Landauer T.K. (1998). The measurement of textual coherence with Latent Semantic Analysis. *Discourse Processes*, 25: 285-307.
- Graesser A., McNamara D.S., Louwerse M. and Cai Z. (2004). Coh-Metrix: Analysis of text on cohesion and language. *Behavioral Research Methods, Instruments, and Computers*, 36: 193-202.
- Granger S., Dagneaux E. and Meunier F. (2002). *International Corpus of Learner English*. Handbook and CD-ROM. Louvain-la-Neuve: Presses universitaires de Louvain.
- Hinkel E. (2001). Matters of cohesion in L2 academic texts. *Applied Language Learning*, 12: 111-132.
- Hinkel E. (2005). Analyses of second language text and what can be learned from them. In Hinkel, E., editor, *Handbook of Research in Second Language Teaching and Learning*, Mahwah (NJ): Erlbaum, pp. 615-628.
- Hu X., Cai Z., Wiemer-Hasting P., Graesser A. and McNamara D.S. (2007). Strengths, limitations, and extensions of LSA. In Landauer, T., McNamara, D., Simon, D. and Kintsch, W., editors, *Handbook of Latent Semantic Analysis*, Mahwah (NJ): Erlbaum, pp. 401-425.
- Johnson P. (1992). Cohesion and coherence in compositions in Malay and English. *RELJ Journal*, 23: 1-17.
- Landauer T.K. and Dumais S.T. (1997). A solution to Plato's problem: the latent semantic analysis theory of acquisition, induction and representation of knowledge. *Psychological Review*, 104: 211-240.
- Landauer T.K., Foltz P.W. and Laham D. (1998). An introduction to Latent Semantic Analysis. *Discourse Processes*, 25: 259-284.
- Landauer T.K., McNamara D., Simon D. and Kintsch W. (2007). *Handbook of Latent Semantic Analysis*. Mahwah (NJ): Erlbaum.
- Landauer T.K. and Psotka J. (2000). Simulating text understanding for educational applications with Latent Semantic Analysis: Introduction to LSA. *Interactive Learning Environments*, 8: 73-86.
- Lebart L., Morineau A. and Piron M. (2000). *Statistique exploratoire multidimensionnelle* (3ième édition). Paris: Dunod.
- Lonsdale D. and Strong-Krause D. (2003). Automated Rating of ESL Essays. In *Proceedings of the NAACL 2003 Workshop*, pp. 61-67.
- Louwerse M.M., McCarthy P.M., McNamara D.S. and Graesser A.C. (2004). Variation in language and cohesion across written and spoken registers. In *Proceedings of the 26th Annual Conference of the Cognitive Science Society*, pp. 843-848.

- 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*. Doctoral dissertation, University of Memphis.
- McNamara D.S., Cai Z. and Louwerse M.M. (2007). Optimizing LSA measures of cohesion. In Landauer, T., McNamara, D., Simon, D. and Kintsch, W. editors, *Handbook of Latent Semantic Analysis*, Mahwah (NJ): Erlbaum, pp. 379-400.
- McNamara D.S., Louwerse M.M., Cai Z. and Graesser A. (2008). Coh-Metrix version 2.1. Retrieved 22.09.2008, from <http://cohmetrix.memphis.edu>.
- McNamara D.S., Ozuru Y., Graesser A.C. and Louwerse M. (2006). Validating Coh-Metrix. In *Proceedings of the 28th Annual Conference of the Cognitive Science Society*, pp. 573-578.
- Mojica L.A. (2006). Reiterations in ESL learners academic papers: Do they contribute to lexical cohesiveness? *The Asia-Pacific Education Research*, 15: 105-125.
- Rehder B., Schreiner M.E., Wolfe M.B.W., Laham D., Landauer T.K. and Kintsch W. (1998). Using latent semantic analysis to assess knowledge: Some technical considerations. *Discourse Processes*, 25: 337-354.
- Scardamalia M. and Bereiter C. (1987). Knowledge telling and knowledge transforming in written composition. In Rosenberg, S., editor, *Advances in applied Psycholinguistic (II): Reading, writing and language learning*, Cambridge: Cambridge University Press, pp.142-175.
- Shapiro A.M. and McNamara D.S. (2000). The use of latent semantic analysis as a tool for the quantitative assessment of understanding and knowledge. *Journal of Educational Computing Research*, 22: 1-36.
- Silva T. (1993). Toward an Understanding of the Distinct Nature of L2 Writing: The ESL Research and Its Implications. *TESOL Quarterly*, 27: 657-677.
- Strömquist S. and Day D. (1993). On the development of narrative structure in child L1 and adult L2 acquisition. *Applied Psycholinguistics*, 14: 135-158.
- Thewissen J. (2008). The phraseological errors of French-, German-, and Spanish-speaking EFL earners: Evidence from an error-tagged learner corpus. In *Proceedings from the 8th Teaching and Language Corpora Conference*, pp. 300-306.
- Thewissen J. (forthcoming). *Accuracy across proficiency levels and L1 backgrounds: Insights from an error-tagged EFL learner corpus*. PhD Thesis. Centre for English Corpus Linguistics. Université catholique de Louvain, Belgium.
- van Bruggen J., Sloep P., van Rosmalen P., Brouns F., Vogten H., Koper R. and Tattersall C. (2004). Latent semantic analysis as a tool for learner positioning in learning networks for lifelong learning. *British Journal of Educational Technology*, 35: 729-738.
- Wiemer-Hastings P. and Graesser A. (2000). Select-a-Kibitzer: A computer tool that gives meaningful feedback on student compositions. *Interactive Learning Environments*, 8: 149-169.

