

## LATENT SEMANTIC ANALYSIS: AN ANALYTICAL TOOL FOR SECOND LANGUAGE WRITING ASSESSMENT

İhsan ÜNALDI

Gaziantep Üniversitesi Yabancı Diller Yüksek Okulu

Doç. Dr. Yasemin KIRKGÖZ

Çukurova Üniversitesi

### Abstract

*This small-scale study tries to introduce and validate a linguistic theory and method, Latent Semantic Analysis. This method has been used in natural language processing for some time now; however, its potential in examining second language lexicon is a relatively recent discussion (Crossley, Salsbury, McCart and McNamara, 2008). In this study, the potential of this method to distinguish between native English text sets and texts written by Turkish university students learning English as a foreign language were analyzed. To test this potential, an original specialized corpus was formed from essays written by these Turkish university students, and it was compared to another corpus involving native English text sets. In this comparison, three levels of semantic indexing were used: sentence, paragraph and all combinations in texts. The results confirmed the findings in the literature claiming LSA's ability to distinguish between native and learner text sets. Prospective potentials of this method in terms of language teaching-learning context are discussed.*

**Key Words:** Latent Semantic Analysis, learner corpus, cohesion

## GİZLİ ANLAM ANALİZİ: İKİNCİ DİL KOMPOZİSYON DEĞERLENDİRME SÜRECİNDE KULLANILABİLECEK ÇÖZÜMSEL BİR ARAÇ

### Özet

*Küçük ölçekli bu çalışma, dilbilim bağlamında Gizli Anlam Analizi kuram ve yönteminin tanıtımını ve geçerlilik sınavasını içermektedir. Bu yöntem, doğal dil işleme süreçlerinde uzun zamandır kullanılmaktadır fakat bu yöntemin yabancı dil kelime kapasitesi açısından ele alınması yeni bir tartışma alanıdır (Crossley, Salsbury, McCart ve McNamara, 2008). Bu çalışmada, bu yöntemin anadili İngilizce olan üniversite öğrencileri tarafından yazılmış metinlerle, anadili Türkçe olan öğrenciler tarafından yazılmış İngilizce metinleri ayırt etme potansiyeli analiz edilmektedir. Bu potansiyelin sınanabilmesi için anadili Türkçe olan öğrenciler tarafından yazılmış İngilizce metinlerden oluşan özel amaçlı bir derlem oluşturulmuş ve bu derlem anadili İngilizce olan üniversite öğrencileri tarafından yazılmış metinleri içeren başka bir derlemle karşılaştırılmıştır. Bu karşılaştırma sürecinde üç çeşit anlam indeksi kullanılmıştır: tümce, paragraf ve metin geneli. Sonuçlar, bu yöntemin literatürde geçen diğer çalışmalarda ortaya çıkarılmış olan yeterliğiyle paralellik göstermektedir. Bu yöntemin, yabancı dil öğretim-öğrenim bağlamındaki olası yeri tartışılmaktadır.*

**Anahtar Kelimeler:** Gizli Anlam Analizi, öğrenci derlemi, anlamsal uyum

## Introduction

This is a small-scale study involving mathematical analysis of English texts written by Turkish university students learning English as a second language. A specialized corpus was constructed from these texts and it was compared to a similar native corpus. In the process, the dependent variable Latent Semantic Analysis (or Latent Semantic Indexing) method was tested to see if it is able to discriminate between these two corpora.

### ***What is Latent Semantic Analysis***

Latent Semantic Analysis (henceforth LSA) is a linguistic theory and method. It has been used in natural language processing to determine semantic relationships in large bodies of corpora. This mathematical technique is mostly used in popular search engines like Google to categorize and organize large bodies of texts.

LSA makes use of Singular Value Decomposition (henceforth SVD), a mathematical matrix decomposition technique which is used to reduce thousands of dimensions and relationships between words to a more manageable number. In this respect, SVD is akin to factor analysis (Landauer et al., 1998:262 ). Basically, what is done is converting words in a sentence, paragraph or passage into numerical values by making use of a mathematical technique.

### ***How does it work?***

LSA can be applied to sentences, small paragraphs or large bodies of digitalized texts. As the first thing in the process, function words (*stop words* in terms of computational linguistics) are eliminated from the context. These are high frequency words like *am, is, are, and, in etc.*, and to a very large extent, they do not change or relate to the content of the text at hand. Proper nouns (words beginning with uppercase) and abbreviations are also eliminated. For example; when the function words and abbreviations in the following paragraph are eliminated,

*Latent Semantic Analysis (henceforth LSA) is a linguistic theory and method. It has been used in natural language processing to determine semantic relationships in large bodies of corpora.*

we have only the following lexical items left:

*latent semantic analysis linguistic theory method used natural language processing determine semantic relationships large bodies corpora*

In the next step, the lexical items at hand are stemmed. It is a simple process in which words are reduced to their *root* forms. These root forms are called *lexemes*. The above items will look like this after stemming:

*latent semantic analysis linguistic theory method use nature  
language process determine relationship large body corpora*

With these lexemes at hand, a matrix system is constructed by putting them into rows. This matrix is called a *term-document matrix*. Each row represents a unique word, and each column represents the context from which the words are extracted. This context could be a sentence, paragraph or a whole passage. A sample *term-document matrix* is introduced in Table 1.

**Table 1:** A Sample Term-document Matrix

	Document 1	Document 2	Document 3	Document n
Lexeme 1	1	2	0	N
Lexeme 2	1	0	3	N
Lexeme 3	1	1	0	N
Lexeme 4	1	0	0	N
Lexeme 5	0	0	4	N
Lexeme 6	1	1	0	N
Lexeme 7	1	0	0	N
Lexeme 8	0	2	1	N
Lexeme 9	1	1	0	N
Lexeme n	n	n	n	N

In Table 1, each row stands for a stemmed lexeme (Lexeme 1, Lexeme 2 etc...), and each column represents the context i.e. the passage or the text. The numerical values in each cell shows how many times a lexeme occurs in a certain document. For example, Lexeme 1 occurs once in Document 1 and twice in Document 2; however it doesn't occur in Document 3 where it gets a null value.

The next step to take is *term weighting*. In order to determine which words occur more than the others, a *local weighting factor* is calculated. The process is rather simple: words appearing many times in a text are given greater weights than words that appear only once.

A *global weighing factor* is also calculated to determine items that occur many times across the document sets. Like local weighting factor, it is a simplistic word-count process. These two calculations are common, but they are not the only techniques in LSA (see Landauer et al., 1998 for details of this calculation step).

After term weighting, only the nonzero values change in the matrix. However, these values hardly mean anything, as they are regarded as raw data. Suppose that what we have at this point is like a big group of people standing in rows right in front of us, and we want to know the dominant color(s) these people have on their clothes. Looking from a direct angle it is impossible to see the people in the back rows, so we have to change our point of view and find the optimum

angle to catch the whole crowd. The rationale is similar with the last step which is Singular Value Decomposition (SVD), and it is used to reduce the semantic dimensions to smaller manageable units. Again, there are other techniques that could be employed at this stage and SDV is among the options. The outcome is generally a score between 0 and 1, and contexts having close scores to 1 are interpreted as having more cohesion than contexts scoring closer to 0.

An interesting point at this step is that the nonexistence of certain lexical items is as important as what is present in the context. In other words, what is important is not the direct relationships among lexemes but how irrelevant one group of lexemes in a text to the other(s). This point of view is claimed to have shed light onto children's inexplicably fast vocabulary acquisition in their first languages. That is, a child's knowledge about vocabulary when reading is claimed to come from the words that are *not* in the text rather than what is available. For example, a typical American seventh grader learns 10-15 words a day. Landauer (1998:274) suggests that about three-fourths of the gain in total comprehension vocabulary that results from reading a paragraph is indirectly inferred knowledge about words not in the paragraph at all (see Landauer & Dumais, 1997:211-240 for a detailed discussion).

#### ***Synonymy and polysemy issues***

LSA is also related to synonymy and polysemy issues. Words that have *nearly* same meanings are called synonyms. The definition might sound neat, but the reality is somewhat different. Take the word *black* as an example. When you look it up in a thesaurus, you are likely to get a list as follows:

*pitch-dark, black-market, shameful, dim, fatal, inglorious, grim, calamitous, blackened, fateful, dark, sinister, opprobrious, bootleg, disgraceful, smutty, contraband, pitch-black, smuggled, ignominious, disastrous, bleak, mordant*

If you are looking for a word to replace *black* in the following sentence, '*Why are some clouds black?*', you could easily replace it with *dark* from the list obtained from the thesaurus. The others will not be appropriate as they all possibly belong to different contexts. In other words, synonyms are, in fact, context-bound. LSI helps overcome this issue by categorizing texts according to their lexical relations.

Polysemy is another concept which LSA can handle again by making use of lexical relations in a given text. Generally, words in any language have more than one meaning; therefore, a one-to-one mapping of words with their meanings is nearly impossible. The word *bug*, for example, means a kind of insect if it occurs in a text about nature. The same word would mean a kind of digital virus in technological contexts while it would mean a kind of listening device in a detective story. This situation is called polysemy. So, two different documents containing the

word *bug* would be categorized mathematically by making use of the other words surrounding it.

### ***How is it related to second language learning?***

Throughout their learning process, learners produce a variety of outputs either spoken or written. This production state, in fact, is a kind of language with its *universals* and certain stages. This language is often referred to as interlanguage or learner language. Analysis of learner language is not a new field of study. Studies focusing on productions of language learners have quite often taken the stance of error analysis; the specific learner language has been compared with the target language. In other words, contrastive analyses of learner language (spoken or written) have traditionally focused on grammatical issues i.e. '*bad language*' (Ellis, 1990:51). Through this paradigm, researchers have tried to come up with ways to describe prominent features of learners' (L2) productions.

In these contrastive analyses, large bodies of learner corpora have been used. As a matter of fact, related literature doesn't lack analyses of learner corpora. (For a comprehensive bibliography check the following website; [http://sites-test.uclouvain.be/cecl/projects/learner\\_corpus\\_bibliography.html](http://sites-test.uclouvain.be/cecl/projects/learner_corpus_bibliography.html).) Some of these studies related to the current one are reviewed below.

The characteristics of learner language has been researched from numerous aspects. In his study, which could be called as a meta-analysis, Silva (1993:657-677), to develop a clear understanding of the nature of L2 writing, examined 72 reports of empirical research comparing L1 and L2 writing. The results suggested that, in general, adult L2 writing is distinct from and less effective than L1 writing. L2 composing appears to be more constrained, more difficult and less effective. L2 writers appeared to be doing less planning and having problems with setting goals, generating and organizing materials. Their transcribing was more laborious, less fluent, and less productive. Reviewing, rereading and reflecting were less, but they revised more. Naturally, they were less fluent and less accurate. In terms of lower level linguistic concerns, L2 writers' texts were stylistically distinct and simpler in structure. Their sentences included more but shorter T units, fewer but longer clauses, more coordination, less subordination, less noun modification, and less passivization. One important point was about the use of cohesive devices. They used more conjunctive and fewer lexical ties, and exhibited less lexical control, variety, and sophistication. These outcomes have been discussed, analyzed, confirmed and reconfirmed by researchers; however, some aspects of learner language have been ignored.

Among these aspects, cohesion in learner language is a potentially fruitful one and LSA is directly related to cohesion in any text. In language teaching-learning context, trying to deal with *cohesion* in learners' texts is like sailing into uncharted waters; traditionally, it lacks attention (Cook, 1989:127). Theoretical

background concerning cohesion is generally attributed to Halliday & Hasan (1976). They use the word *text* to refer to any passage, spoken or written, of any length, that forms a unified whole. It is a semantic unit which, unlike what is generally accepted, does not consist of sentences, but is *realized by* sentences. When talking about ties, they refer to single instances of cohesion or cohesively related items; and at this point cohesion is considered to have two facets: grammatical and lexical (see Halliday & Hasan, 1976 for details).

This dichotomous view of cohesion, however, has been subject to criticism. Mahlberg claims that cohesion is a fundamentally lexical phenomenon (in Flowerdew and Mahlberg, 2006:118). Pronouns, articles and other function words might, to some extent, help cohesion, but the real ties or cohesion lie in lexical relationships. This point of view goes back to Lewis (1993:91) who states that “(l)anguage consists of grammaticalised lexis, not lexicalized grammar.” When this issue is considered in language teaching-learning context, since the focus is nearly always on the grammar of the target language, the productions of the students are likewise. With the same paradigm, students very often try to produce grammatically correct sentences. So in this respect, it is a bigger surprise to see more cohesive learner essays than grammatically well-crafted ones.

One weak (or to some, strong) point in grammar-focused argument is that, grammar is always neat and systematic. It is easy to notice grammatical problems in a context and give feedback about them; the outcome is rarely disputable. That is, two language teachers trying to grade a learner’s text in terms of grammar will probably come up with similar results. However, when it comes to grading cohesion in the same text (if cohesion is an issue of course), totally different and subjective thoughts might surface. In other words, determining cohesion in a text is a subjective issue.

The concern of the current study is lexical features of L2 writing and the potential of LSA in language teaching-learning is the core aspect. There are certain studies claiming that by making use of LSA, it is feasible to analyze lexical structures of learners’ texts. For example, to determine how computational modeling of semantic knowledge can measure lexical growth of second language learners, Crossley et al. (2008:136-141) carried out a longitudinal study. A small group of L2 learners’ lexical growth was tracked by making use of LSA over a long period of time. The participants were at the lowest proficiency level at the beginning. A spoken corpus was formed through interviews over one year. The data collected in the 2<sup>nd</sup>, 4<sup>th</sup>, 16<sup>th</sup>, 32<sup>nd</sup>, 50<sup>th</sup> and 52<sup>nd</sup> weeks were computed. The results revealed that the values computed in the last meeting (52<sup>nd</sup> week) were statistically significant from those of the first meeting. It was concluded that, in time subjects’ proficiency levels increased in terms of the lexical relations in their utterances.

In another recent and comprehensive study, Crossley and McNamara (2009) explored how lexical differences related to cohesion and lexical networks can be used to distinguish between texts written by first language writers of English and

second language writers of English. Two corpora were used; one was from LOCNESS (Louvain Corpus of Native English Essays), and the other was created from Spanish learners of English. The native corpus comprised of 208 texts (151,046 words in total) and the learner corpus was comprised of 195 essays (124,176). These texts were compared in terms of word hypernymy, word polysemy, argument overlap, motion verbs, CELEX written frequency, age of acquisition, locational nouns, LSA givenness, word meaningfulness, and incidence of casual verbs. These parameters are reported to have distinguished between L1 and L2 texts. The focus being on the syntax and grammar, in the related literature it had been concluded that approaches relying on lexical variables alone could not be used to discriminate L1 and L2 texts (Connor, 1984: 301–316; Reynolds, 1995:185–209). The importance of Crossley and McNamara's study is that it was the first study in which L1 and L2 texts were distinguished solely based on lexical features.

Bearing all the points mentioned above, this study tries to answer the following research question: Can LSA distinguish the texts written by native speakers and the texts written by Turkish students learning English as a second language?

### **Methodology**

This study is a corpus based contrastive analysis of learner language. The participants were freshman engineering students at University of Gaziantep. Before becoming freshman students, all of the participants took English preparation classes for a year. Their ages varied from 19 to 23, and most of the participants were male.

In order to meet the requirements for a learner corpus (Granger, 2003:539), subjects' proficiency levels were determined using a valid and reliable placement test (Allen, 1992). The results were checked to see if their levels were homogeneous as would be expected. However, the results of the test showed that the subjects' levels varied from A2 (elementary) to C2 (advanced) level, which, in our case, demanded adjustments concerning homogeneity. Therefore, only intermediate and upper level subjects (49) were involved in the study. The assumption was that the subjects who had proficiency levels lower than intermediate level would still be dealing with some basic grammar and lexical issues, which was likely to affect the results of this study negatively. From this respect, a purposive sampling strategy is adopted in this study.

The students received training about academic essay writing for about 14 weeks. At the end of the training, students were required to write argumentative essays in a written exam. All the topics in the written exam were taken from LOCNESS (Louvain *Corpus* of Native Speaker Essays, Granger 1998). After the exam, the texts were digitalized by the students themselves. Since the focus is merely on

lexical structures of the learners' texts, during digitalization process students were required to correct their own spelling mistakes. Otherwise the results would have faltered, because misspelled lexical items would have been eliminated at the stemming stage mentioned above. The corrected texts were processed for LSA via an online database, *coh-matrix*. The same procedure was carried out for L1 text sets. Table 2 is a description of the corpora compared in this study.

**Table 2:** *Comparison of Two Corpora Used in the Study*

Name of the Corpus	Total Number of Words	Average Words per Essay	Essay Type	Prompt
Native corpus (L1)	21,605	-500 (400)	Argumentative	Exam/Timed
Learner corpus (L2)	16,334	-500 (333)	Argumentative	Exam/Timed

As it is clear from Table 2, two corpora (L1 and L2) are as similar as possible. The native corpus contains 21,605 words in total with a 400 words of average per essay; and the learner corpus contains 16,334 words in total with an average of 333 words per essay. Both corpora contain argumentative essays which were written by the students during timed exams. Native corpus was taken from Granger (2003), and the learner corpus was originally created by the researchers from the participants' written productions mentioned previously.

The texts from these two corpora were processed through *coh-matrix* one by one. It took about 1-2 minutes for each text to be rendered in *coh-matrix*. The outcomes were in MS Office Excel format. After some minor editing, results for each text were transformed into SPSS whereby group means were compared through t-test.

### Findings

LSA, in *coh-matrix*, is calculated at three levels: sentence, paragraph and all combinations of LSA in a text. The two corpora described in the previous section were analyzed through t-test by taking LSA scores into account. T-test results of Learner (L1) and Native (L2) for LSA at sentence level are presented in Table 3.

**Table 3:** *T-test Results of L1 and L2 Texts in Terms of LSA Sentence Level*

	Levene's Test			$\bar{X}$	sd	df	t	P
	F	Sig.	N					
<b>Learner</b>			49	,184	,055	101		
<b>Native</b>	,001	,979	54	,230	,052	98,611	-4,251	,000



Before carrying out a t-test, homogeneity of variance was checked through Levene’s test of homogeneity of variance. The result of Levene’s test indicates that the population variances are equal ( $p = ,979 > ,05$ ). When it comes to averages, the learner group obtained a mean score of ,184 which is lower than the native group (,230). This difference between groups is statistically significant ( $p < ,05$ ). One could conclude from this result that adjacent sentences in learners’ texts seem to lack cohesion when compared to the native ones.

In coh-matrix, LSA is also computed at paragraph level. The scores obtained from paragraphs in a text are taken into consideration. These scores measure how similar paragraphs are to the other paragraphs in the same text. Obviously, the texts must have more than one paragraph for valid scores. Table 4 introduces LSA t-test results at paragraph level for learner and native text sets.

**Table 4:** *T-test Results of L1 and L2 Texts in Terms of LSA Paragraph Level*

	Levene's Test		N	$\bar{X}$	sd	df	t	P
	F	Sig.						
<b>Learner</b>			49	,386	,096	101		
<b>Native</b>	,126	,724	54	,395	,096	100,048	-,487	,627

As it was explained previously, homogeneity of variance was computed as the first step. The result of Levene’s test in Table 4 indicates that the population variances are again equal ( $p = ,724 > ,05$ ). However, when mean scores of the groups are taken into consideration, there seems to be a slight difference between the two groups (the learner group = ,386 and the native group = ,395). This slight difference doesn’t appear to be statistically significant ( $p = ,627 > ,05$ ).

The third LSA level in coh-matrix is combinations of all sentences. This computation takes all sentence combinations into account, not just adjacent sentences. T-test results concerning all sentence combinations are exhibited in Table 5.

**Table 5:** *T-test Results of L1 and L2 Texts in Terms of LSA All Combinations*

	Levene's Test		N	$\bar{X}$	sd	df	t	P
	F	Sig.						
<b>Learner</b>			49	,166	,052	101		
<b>Native</b>	,984	,324	54	,214	,062	100,291	-4,201	,000

Levene’s test result presented in Table 5 reveals that the population variances are equal ( $p = ,324 > ,05$ ). Learner group mean (,166) is lower than the native group (,214); and this difference appears to be statistically significant ( $p = ,000 < ,05$ ). Since the mean score of the native group is significantly higher than the

learner group, it wouldn't be a presumption to claim that the learner group's texts lack the coherence that the native group's texts have.

Statistical analyses of learner and native groups indicate that there are significant differences between the two groups in terms of LSA at sentence and all combinations level. However, when the problem at hand is cohesion, numbers only hardly mean anything. In order to make the point about cohesion clear, the following extracts from four different essays written by learners involved in the study are exhibited. The possible causes of the significant differences between the learner and the native text sets might be seen clearly in these extracts.

Extract 1:

*There are a lot of brochures, newspapers at the newspaper markets. Every brochure cannot be good, because sometimes there are magazines brochures between them. The brochures usually are read by young people.*

Extract 2:

*A real or theoretical world in university! How do we grow up? Actually, a number of students always think that they are different from other people. Most of universities just give information as theoretic. Thus, I believe that real world is more important than theoretical world for realizing prospective of things that they will use.*

Extract 3

*Gun power is getting so important all over the world where the world smells like a battle area. Almost every single technology is involved in the war industry. New technology war toys, they are not called as guns anymore, can cause catastrophic results. However, they are nothing without a thinking brain. Hence, the question is that, should the kids who are going to play with these toys be professionals or anyone?*

Extract 4

*If you wanted to harm a nation, it would be enough that you should take their media. TV, radio, newspapers are the most important things, and the easiest way for reaching all the people. If the media is healthy, the nation is healthy too, and in Turkey, we are not healthy.*

In the extracts above, written by four different learners, cohesion is undermined by a lack of ties between sentences, which makes the sentences look like floating around in a disconnected manner; so LSA results seem to be consistent with what can be observed from learners' essays.

### Discussions

The findings in this study are in line with related literature Crossley et al. (2008: 136-141) and Crossley and McNamara (2009: 119-135). LSA scores of native speakers of English at sentence level and all combinations are statistically different from the scores obtained from learners of English. However, scores at paragraph level showed no significant difference.

This study could be counted as a validation of LSA with Turkish learners of English as a second language. When it comes to the application of LSA in language teaching-learning context, assessment of second language writing is an open door for LSA into the domain.

There are three major approaches to scoring writing performances of second language learners. These approaches are: holistic, primary trait and analytical (Brown, 2003:241). In holistic scoring, the evaluators or the raters follow prescribed patterns. There are holistic scales where each point is given a systematic set of descriptors and the evaluator tries to arrive at a score by matching overall impressions with these descriptors. In primary trait scoring, the objective of written task is taken into consideration. If, for example, the objective of a written task is to write about a certain process, the score goes up as the learner gets close to achieving the mentioned task. In the last scoring system, analytical scoring, the written production is evaluated through analytical parameters like organization, grammar and lexicon, and some parameters might be added to or subtracted from these parameters in accordance with teaching-learning objectives.

Automated essay analysis systems have been used since 1960's, and a number of software has been developed for holistic scoring (see Burstein and Chodorow, 2002:487-497 for details). Discourse, syntactic and content analysis are among the challenges to be faced for these systems. The main challenge, however, is to come up with a single holistic *machine* score totaled from analytical scores like syntactic or lexical. LSA seems to be a potential parameter in this kind of analytical scoring.

Many language teachers would be tempted with the idea of computer software grading learners' writings. Our professional lives would really be less demanding if we had such software grading our students' essays while we are sipping our coffee. However, as the proverb suggests, "If it sounds too good to be true, it probably is." Rather than a clear-cut categorizing like human raters vs. machine raters, these two aspects should be regarded as complementation of each other. LSA scores, along with other analytical and numerical values could be complementation to human raters during second language writing assessments.

### References

- Allen, D. (1992). *Oxford placement test 2 (New edition)*. Oxford: Oxford University Press.
- Aston, G., Bernardini, S. and Stewart, D. (2004). *Corpora and Language Learners*. Amsterdam / Philadelphia: Benjamins.
- Brown, D. H. (2003). *Language Assessment - Principles and Classroom Practice*. New York: Longman.
- Burstein, J., Chodorow M. (2002). "Directions in Automated Essay Scoring." In *Handbook of Applied Linguistics*, pp. 487–497. Oxford: Oxford University Press.
- Connor, U. (1984). "A study of cohesion and coherence in ESL students' writing." *Papers in Linguistics: International Journal of Human Communication*, 17, 301–316.
- Cook, G. (1989). *Discourse*. Oxford: Oxford University Press.
- Crossley, S. A., Salisbury, T. McCarthy, P. & McNamara, D. S. (2008). "Using Latent Semantic Analysis to explore second language lexical development." In D. Wilson & G. Sutcliffe (Eds.), *Proceedings of the 21st International Florida Artificial Intelligence Research Society* (pp. 136-141). Menlo Park, California: AAAI Press.
- Crossley, S. A., McNamara, D. S. (2009). *Computational assessment of lexical differences in L1 and L2 writing*. *Journal of Second Language Writing* (June 2009), 18 (2), 119-135
- Ellis, R. (1990). *Instructed Second Language Acquisition*. Oxford: Basil Blackwell.
- Granger, S. (ed.) (1998). *Learner English on Computer*. London & New York: Addison Wesley Longman
- Granger, S. (2003.) *The international corpus of learner English: a new resource for foreign language learning and teaching and second language acquisition research*. *TESOL Quarterly* 37(3): 538-545.
- Flowerdew, J. and M. Mahlberg (eds.) (2006). *Corpus Linguistics and Lexical Cohesion*. Special issue of the *International Journal of Corpus Linguistics*.
- Halliday, M. A. K., and Hasan, R. (1976). *Cohesion in English*. London: Longman.
- Landauer, T. K., and Dumais, S. T. (1997). *A solution to Plato's problem: The Latent Semantic Analysis theory of the acquisition, induction, and representation of knowledge*. *Psychological Review*, 104, 211-240.
- Landauer, T. K., Foltz, P. W., & Laham, D. (1998). *Introduction to Latent Semantic Analysis*. *Discourse Processes*, 25, 259-284.
- Lewis, M. (1993). *The Lexical Approach. The State of ELT and a Way Forward*. Hove: Language Teaching Publications.
- Reynolds, D. W. (1995). *Repetition in nonnative speaker writing*. *Studies in Second Language Acquisition*, 17, 185–209.
- Silva, T. (1993). *Toward an understanding of the distinct nature of L2 writing: The ESL research and its implications*. *TESOL Quarterly* 27, 4:657-77.