Lexical errors in second language scientific writing: Some conceptual implications

MARÍA LUISA CARRIÓ-PASTOR* & EVA MARÍA MESTRE-MESTRE
Universidad Politécnica de Valencia

ABSTRACT
Nowadays, scientific writers are required not only a thorough knowledge of their subject field, but also a sound command of English as a lingua franca. In this paper, the lexical errors produced in scientific texts written in English by non-native researchers are identified to propose a classification of the categories they contain. This study will enable researchers to improve their writing and facilitate smoother communication among international writers. In addition, establishing the causes of these errors may enable the recurrent pattern to be identified and the necessary guidelines for their correction to be drawn up. These data may be able to illuminate the processes followed by non-native speakers of English when learning new words, and thereby facilitate the avoidance of errors and the identification of the mechanisms which can permit the correct production of the specialised lexicon.

KEYWORDS: error analysis; error classification; second language acquisition; scientific writing.

RESUMEN
En la actualidad, los escritores científicos han de ser no sólo conocedores de sus areas específicas de conocimiento, sino también de la lengua inglesa, que se utiliza como lengua franca. En este artículo, se han identificado los errores léxicos que se producen en los textos científicos escritos en inglés por investigadores no nativos para proponer una clasificación de sus categorías. Este estudio permitirá a los investigadores mejorar su escritura y facilitará una mayor comunicación entre los escritores internacionales. Adicionalmente, el establecer las causas de estos errores podría permitir identificar los patrones recurrentes y proponer una serie de medidas para corregirlos. Estos datos podrían mostrar los procesos que han seguido los escritores no nativos de la lengua inglesa cuando aprenden nuevas palabras y, por lo tanto, facilitar no cometer errores e identificar los mecanismos que pueden permitir una correcta producción de un léxico especializado.

PALABRAS CLAVE: análisis de errores; clasificación de errores; adquisición de una segunda lengua; escritura científica.

_____
Address for correspondence: Mª Luisa Carrió-Pastor. Departamento de Lingüística Aplicada, Universitat Politècnica de València, Camino de Vera s/n, 46022 Valencia, Spain; e-mail: lcarrio@idm.upv.es.

1. INTRODUCTION

Due to its intrinsic nature, scientific language has a series of implicit features which render it completely different from any other genre. It possesses a series of characteristics which are inherent to scientific thought and expression. Scientific papers are *recognisable*, in terms of wording as well as structures. Scientific language can be said to be a conceptual map of some kind, in which readers *jump* from one marked concept to another. In a scientific text, an experienced reader can follow the written pathway laid out in front of him or her, which is full of *common landmarks* (concepts). This also hinders its understanding and production by the layperson, with a range of specialised terms being employed, which vary depending on the audience being addressed.

In this respect, Gotti (2003) describes several lexical and grammatical features which characterise scientific language: *extremely compact* syntactic structures, the omission of articles or prepositions in order to obtain conciseness, the avoidance of relative clauses and subordination, complex pre-modification and nominalization (for the purposes of precision and depersonalization).

Alcaraz Varó (2000: 138-9) also makes reference to the nature of scientific-technical nouns in English when the means of communication is the research article: "[…] la alta densidad sémica o conceptual de las unidades léxicas compuestas; el empeño por la precisión expresiva, materializado en los sintagmas nominales largos".

Due consideration must be given to the cognitive, linguistic and socio-communicative components when analysing the effects of scientific writing, particularly with regard to the use of lexis. With regard to lexis, the abovementioned features can be seen in the high density of specialised lexicon in scientific texts. The meanings of precise or specialised lexical units are not explained by authors when used in scientific texts; thus it is presumed that the reader is already familiar with them. It can even be said that they are a kind of code shared by both reader and writer. In another line, Arden-Close (1993) and Mudraya (2006) underlined the importance of the lexicon in the acquisition of a second language, since it is a source of error. They indicate that more attention should be paid to it, and that meaning should be explained from a lexical perspective.

Corder (1967) and Richards (1971; 1974) studied error analysis to understand language behaviour. Its main aim was second language acquisition, but researchers were also interested in understanding the linguistic aspects of error production. Some decades later, interest in the concept of linguistic relativity with regard to error production has now been rekindled. The theory here proposed is as follows: if speakers of different languages do not understand each other, the reason is not that their languages do not lend themselves to translation (which they obviously do), but that they observe and interpret reality in significantly different ways. In this sense, error production may be avoided if the ways to communicate in a second language could be explained.

The concepts which words signify may not be represented in the same way; that is, the understanding of another language does not depend on identifying structures which are equivalent to those of the mother tongue but on equivalence between the concepts emerging from reality and then identifying the appropriate way of expressing these. As Yoshii (2006: 88) remarks, "[…] L2 learners rely on word-to-word links in early stages, but as their L2 proficiency develops, they link L2 directly to concepts (conceptual links)". This is directly related to the implied mono-referentiality of nouns used in science. We should consider that nouns help us to communicate and express our thoughts, and consequently change with these, hence preventing total equivalence between languages. In this sense, in order to unravel the mechanisms behind language, a cognitive interpretation of language becomes essential.

Error Analysis has helped in the understanding of error not merely as an unwanted phenomenon in language, but as a source of information which can be used to improve production in a second language. The errors found in writing can illuminate the writing process and help us to understand the mechanisms that the non-native speaker adopts. As a result, by understanding these error patterns, several strategies may be designed to improve writing in a second language and several different issues such as the cognitive processes of language production could be considered when analysing errors.

The first important issue in this paper is error identification. The correct identification of errors serves to establish the causes and the processes followed in language production. Many studies concerning errors have focused on the nature of these, but very few have analysed the ability to identify and interpret errors in a second language (Rifkin & Roberts, 1995; Carrió Pastor, 2004; Hamid, 2007; Mestre, 2011). A further issue of importance in this research is that second language errors are a result of different causes. Traditionally, these causes can be divided into two categories: *interlingual errors*, which are due to first language interference upon the second language and *intralingual errors*, which are produced regardless of the mother tongue and are due to deficiencies in the learning process (James, 1998: 179; Larsen-Freeman & Long, 1992: 58). To this well-known classification, we add a third: *conceptual errors*, caused by the failure of the speaker to match an idea with the correct expression, i.e. a breakdown of the concept-term relationship. We consider that apart from the errors caused by the interference of the mother tongue or the deficiencies in second language acquisition which might refer to linguistic and socio-communicative aspects, a third cause of errors can be due to cognitive aspects of language production, i.e. conceptual interference. This third cause could explain several errors that are caused by the erroneous conception of the relationship between image, concept and term. Speakers of a second language should be conscious of the fact that words are only representations of concepts and they should learn how to associate one concept to several terms if they speak several languages. This fact is clear enough when we refer to synonyms in our mother tongue, and yet this association is not systematically applied when acquiring a second language.

The second issue in this paper is related to lexical errors. Webber (1993) states that the most common causes of errors in non-native English speakers are lexical in nature, due to mother tongue interference. Further underlining the importance of lexicon, more recent studies suggest that lexical but also grammar structures are the most difficult aspects to correctly reproduce in a second language in the different stages of language acquisition (Al-Jarf, 2000; Carrió & Seiz, 2000; Levinson, Lessard & Walter, 2000; Carrió, 2004; Carrió, 2009; Carrió & Mestre, 2010). As well as acknowledging that the study of lexical errors is particularly prolific in the analysis of second language acquisition, these studies agree that errors should no longer be regarded negatively, but rather as an opportunity for improvement (Carrió, 2004).

We believe that the compilation of a corpus of lexical errors could facilitate the understanding of conceptual implications in second language acquisition, student progression and development and also course and material design (Hunston & Francis, 2000; Belz, 2004; Chapelle, 2004; Nelson, 2006; Krishnamurthy & Kosem, 2007). Thus, the compilation of a corpus of lexical errors could help to determine why the concept whose communication is attempted is not universal and which, therefore, depends on cultural conceptions. Language constitutes evidence of the multiple conceptions of reality and these conceptions are expressed through the filter of convictions, culture and linguistic conventions. This diversity is greater still in multilingual contexts.

The third important issue in this paper is error classification and causes. Lexical errors have traditionally been classified according to formal, vocabulary-related considerations or from a semantic perspective. The most well known formal classification of lexical errors (James, 1998: 145) is: *mis-selection* (wrong word choice), *misformation* (words that are non-existent in the L2 but exist in L1) and *distortion* (words that are non-existent in both the L2 and the L1). With regard to semantic errors in lexis, there are two main types: *confusion of sense relations* (a word being used in contexts where a similar word should be used) and *collocational errors* (the choice of a word to accompany another is inappropriate). The interest in classification for this present study derives from the need to establish the causes of the errors produced in an L2 by scientific researchers. By doing this, the factor which causes them can be determined, and we may know whether this entails cognitive causes.

In this sense, the objectives of this study are: firstly, to elaborate a corpus of specialised lexical errors that appear in scientific texts produced by non-native English writers; secondly, to identify lexical errors and their most significant causes in order to generate guidelines which can help improve written production; finally, as a result of the aforementioned objectives, to propose a new classification of lexical errors, including the conceptual component in lexical production that sometimes causes the inappropriate relations established by the second language writer when making the error.

2. METHODOLOGY

From the outset, the type of corpus used in the study and the final conclusions upon the lexical errors made in scientific English was established. The corpus of the research was provided by the Proof-reading and Translation Service of the Universitat Politècnica de València. It consists of thirty scientific papers written by researchers belonging to this university and the same thirty papers corrected by native English proof-readers. The field of these papers was engineering. All the information related to the authors was eliminated from the corpus and only some parts of the texts were used as examples in this research.

In this study, the first set of papers is referred to as the *original papers* and the second group of papers is referred to as the *corrected papers*. The *original papers* were written by Spanish researchers holding an upper-intermediate (B2) level of English, according to the European Common Framework of Reference for Languages. They were intended for publication in international journals, and had thus been sent to the proof-reading service offered by the university in order to ensure that the English was of the appropriate standard.

Once the papers had been selected, all tables, graphs and references were removed and the documents were saved in a text format, in order to enable the data to be analysed using the software *Wordsmith Tools 5.0* (Scott, 2009). Next, the aligning tool in the *Wordsmith Tools* software was used to identify the errors, by comparing each original sentence with the same sentence after proof-reading. Once the errors had been identified, they were classified according to the different categories explained above. The coding process used was through the tagging of the different errors. Six raters were involved in this research and tagged lexical errors. We designed a grid (see Table 1) with the different categories found of the interlingual, intralingual and conceptual errors in order to guarantee reliability of the coding of the corpus:

INTERLINGUAL ERRORS	INTRALINGUAL ERRORS	CONCEPTUAL ERRORS
Calques	Erroneous collocation	Use of a word due to confusion over meaning
Adaptation of words from L1 to L2	Coinages	Error due to confusion of form
Unnecessary borrowings	Omission of part of words	Use of a general word instead of a specific word
	Misformation of words	Wrong near-synonym
	Misordering of words	

Table 1. Classification of errors

Instances of errors can be seen in Examples 1, 2 and 3:

[Example 1. Interlingual errors]
Calques:
> In this work, microwave heating of rubber compositions was *realized/performed* in a modified cylindrical cavity
> These lower temperatures of the rubber, mainly located at the edges of the sample, are *traduced/translated* into a considerably […]

Adaptation of words from L1 to L2:
> Scans for the (006) peak indicate a c-axis spread *of the order of/in the range of* 1°. This method allows solving the heat equation for *bidimensional-meshed domains/two-dimensional meshed*

[Example 2. Intralingual errors]
Erroneous collocation:

A linear equation for [...] has been shown to *reproduce well/accurately* the experimental results

Note that for *zero/no* penetration, we will have infinite parallel and null serial capacitance

Coinage:

three typlets/ triplets

nonprofit/ not-for-profit

Omission:

scale of *valuation/evaluation*

[Example 3. Conceptual errors]

The use of a word instead of another due to confusion over meaning:

Both points of views are considered to be *related/connected*

This method *allows solving/the resolution* of the heat equation

Words that are formally similar:

A sensibility analysis is *processed/produced*

Has to know the *reliability degree/how reliable*

The use of a general word instead of a more specific word:

We have developed a *fabrication/manufacturing* process

the destruction of phenolic wastes has been *tried/tested* on bench and pilot plant scale

tried/ tested*; specific/* particular*; happens/* takes place*; stay still/* remain*; use/* employ*; direction/* path

Selection of inappropriate near-synonym:

[It] is possible [to] include a *grill/grid* too

This step must be carried out *conveniently/carefully* so as to not influence the final result

Raters discussed some borderline cases in order to classify errors in the same categories. Finally, the frequency of the occurrences of errors was calculated in order to obtain information about its importance in the results. The different frequencies were then compared and the causes were drawn from the analysis.

3. RESULTS

The corpus used in this paper for the detection and categorisation of errors displays the features shown in Table 2 below. It can be seen that the number of sentences, paragraphs, words and lists of words diminishes in the texts corrected by the native proof-readers of the papers. The texts were shortened during the revision process, seeming to imply that the Spanish authors used more words than necessary to express themselves in English language. The original texts therefore demonstrate a divergence from the conciseness of expression required in technical English.

As explained in the Introduction, the errors found in the corpus were classified according to their underlying causes as *interlingual errors* or *intralingual errors*, following James (1998) and Larsen-Freeman & Long (1992), with a further type also being distinguished, namely *conceptual errors*. As explained above, some errors may be caused by the misinterpretation of concepts in the target language. The process followed to translate a concept from the mother tongue taking into the target language is not always followed correctly by speakers of a second language and so conceptual errors may be detected.

STATISTICAL DATA		ORIGINAL PAPERS	CORRECTED PAPERS
	Words in the corpus	110,154	108,535
	List of words	8,110	7,583
	Number of sentences	5,468	5,416
	Average number of words per sentence	20.1	20
	Number of paragraphs	1,755	1,701
	Number of sentences per paragraph	3.1	3.2

Table 2. Statistical data of the articles integrating the corpus of original and corrected texts

Subsequently, we sub-divided these groups into categories depending on the type of error. Exhaustive knowledge of the causes of errors provides relevant information to the linguist, in that it can help to determine the relations and conceptual associations established by the non-native speakers. This information can also be of use to L2 teachers, since these results can highlight those aspects which need reinforcement during the learning process.

The first group of errors is that of interlingual errors or interferences, which arise due to L1 interference (Spanish), since the sentence structure, as well as word formation, present a pattern based on the mother tongue. In this group we found the categories of calques, the conversion of words from the mother tongue into words that do not exist in the target language and borrowings. Table 3 shows the errors obtained caused by the interlingual interference:

	CLASSIFICATION OF ERRORS	OCCURRENCES (%)
INTERLINGUAL ERRORS	Calques	53 (65.4%)
	Adaptation of words from L1 to L2	15 (18.5%)
	Unnecessary borrowings	13 (16.1%)
	Total	81 (100.0%)

Table 3. Errors caused by interlingual interference

The most significant datum is that more than half the errors consist of the use of linguistic calques (65.4%) that is L1 greatly influenced the choice of vocabulary used in the texts. However, word adaptations from L1, which do not exist in L2, are scarce and therefore were not statistically relevant in the global result of errors due to L1 influence.

The second group comprises the intralingual errors or interferences, caused by generalizations based on partial exposure to the target language. Second language learners try to generate the rules which govern the data to which they have been exposed, and may develop hypotheses that correspond neither to the mother tongue nor to the target language. In this group we found errors related to the omission of parts of the words, the misformation of words, misordering of words, the erroneous collocations of words and coinages (invention of new words applying erroneous rules of L1). In Table 4 we can observe the occurrences and percentages found:

INTRALINGUAL ERRORS	CLASSIFICATION OF ERRORS	OCCURRENCES (%)
	Erroneous collocation	32 (53.3%)
	Coinages	12 (20.0%)
	Omission of part of the word	11 (18.3%)
	Misformation of words	3 (5.0%)
	Misordering of words	2 (3.4%)
	Total	60 (100.00%)

Table 4. Errors caused by intralingual interference

Table 4 shows that the most frequent types of errors in this group were erroneous collocation and coinages, which account for more than 70% of errors. Omission of some part of the word comes next, although these errors were not numerous in terms of the overall results. The other two sources of errors had insignificant occurrences, as demonstrated by the percentages.

Finally, the third group consists of those errors that arise due to confusion between concept and term, with the results being displayed in Table 5. As can be observed, these errors are the most numerous in the corpus:

CONCEPTUAL ERRORS	CLASSIFICATION OF ERRORS	OCCURRENCES (%)
	Use of a word due to confusion over meaning	53 (12.4%)
	Error due to confusion of form	102 (23.4%)
	Use of a general word instead of a specific word	120 (27.4%)
	Wrong near-synonym	161 (36.8%)
	Total	436 (100.0%)

Table 5. Conceptual errors

The erroneous choice of a word due to confusion over its meaning is not due to L1 influence. It arises because the Spanish writer associates the word with the literal meaning, choosing one over the other because of their similar forms. That is, in the case of choice between two near-synonyms, the Spanish writers chose a word because they misinterpreted its meaning. The most frequent errors were produced as a result of having selected a general word significant but inappropriate for the context of the paper instead of a specific term. The second most common type of errors found was erroneous collocation.

Figure 1 displays the results obtained from the analysis of the three groups of errors made by Spanish researchers when writing research papers in English. As we can see, most of the errors were caused by conceptual interferences: the Spanish writers with a B2 level of English were unable to link the correct word with the right concept, choosing an inappropriate expression – yet this was not caused by a grammatical failure or the influence of the mother tongue. Words with a similar meaning confused the writers, preventing them from choosing the correct option.
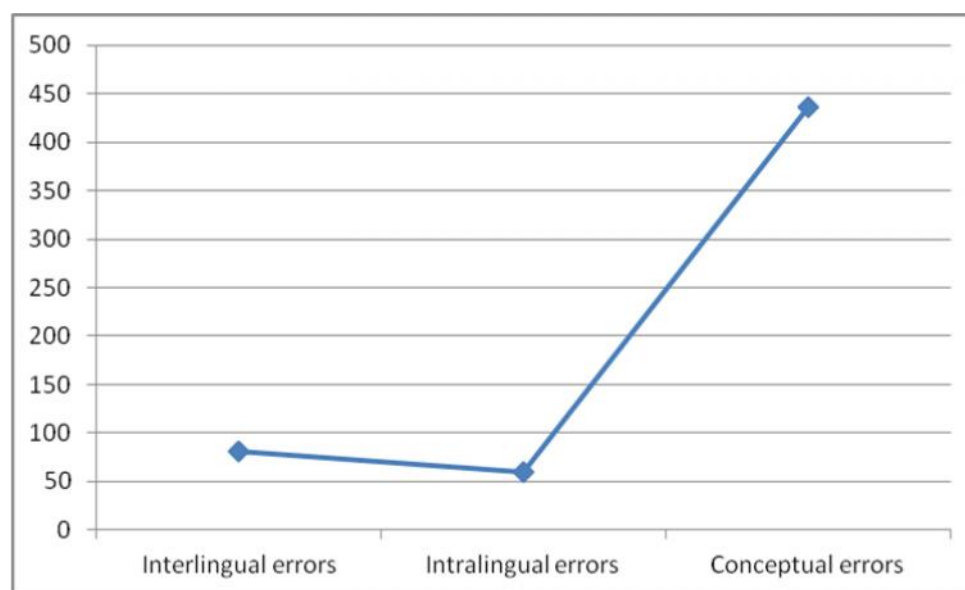
Figure 1. Results of the interlingual, intralingual and conceptual errors

4. CONCLUSIONS

The initial aim, which provided the motivation for this study, was the classification of errors produced by Spanish writers when using English as a lingua franca for the publication of research in international journals. Traditionally, errors have been divided into grammatical or semantic errors, taking a linguistic perspective, or into *interlingual* or *intralingual* errors, from a didactic perspective, taking into account the influence of the language learning process or of the mother tongue. However, in this study another type of error has been studied, introducing a cognitive classification: conceptual errors.

We would like to highlight that the existing relationship between object, concept and term is not universal and unalterable in language. Some concepts have several forms of representation, which vary depending on the cultural background of the speakers, who select one term rather than another according to their specialised knowledge of the subject matter, the socio-communicative components inherent in each act of communication and the learning processes they may have experienced in second language acquisition. The fact that certain errors exist due to the influence of the mother tongue, to lexical distortion or incorrect spelling, to the erroneous choice of a term or to an inadequate conceptual association can be adduced as evidence for the position that these concepts are not associated to specific terms, and that there is no universal form by which these can be labelled.

As detailed in the results section of this paper, conceptual errors were the most prolific cause of error in the English-language texts analysed in this study; in particular, the most frequent cause of error was the subcategory of the choice of a word with a similar meaning to another. This shows that non-native speakers of English with an upper-intermediate (B2) level of English had problems finding the conceptual equivalence between terms and objects despite the fact that their grammatical proficiency is sufficiently adequate to enable

communication. That is to say, if a term is learnt in a language, with certain specifications, and implications, to what extent it is completely translatable to an L2. These results led to the following questions: what exactly is the process that helps us relate equivalent terms in two languages? Can we completely equate terms, which refer to concepts or ideas in two different languages? The teaching of a second language is at present carried out using a communicative approach. Concepts, and not terms, are taught with this method. This may be the reason for the errors found in the corpus. The Spanish writers know conceptually the specific term in their L1. The problem arises when this concept has to be associated to the L2. Therefore, the process would be: object-concept-two terms (L1/L2). According to our results, the writer with an upper-intermediate (B2) level of language proficiency has not absorbed this process, and simply relates an object or concept to the L1 term and then translates it to the L2 term. The teaching strategies for foreign languages should transmit that the relationship between concepts and terms is multi-faceted, and is not unidirectional, but travels in as many directions as the languages known to the speaker.

Further study of the relationship between object-concept-term, since a clear understanding of this is vital for an error-free production in a second language. Due attention during teaching and learning must be paid not only to the linguistic processes, but also the cognitive and socio-communicative processes of the speakers of a second language in order to ensure the accurate expression of ideas.

It has not been our intention to draw conclusions regarding error correction, which can be found in articles purely dedicated to didactics, and which focus on the learning of a foreign language (Gaskell & Cobb, 2004; Lee, 2004; Salem, 2007). However, we are well aware that the conclusions of this study could be applied to the field of foreign language teaching, both from the point of view of the design of learning strategies and from the creation of materials aimed at practising and correcting the aspects which most frequently elicit errors from non-native writers of the English language.

REFERENCES

Al- Jarf, R. (2000). Grammatical agreement errors in L1/L2 translations. *IRAL, 38*, 1–15.

Alcaraz Varó, E. (2000). *El inglés profesional y académico*. Madrid: Alianza.

Arden-Close, C. (1993). Language problems in science lectures to non-native speakers. *English for Specific Purposes*, *12*(3), 251–261.

Belz, J. A. (2004). Learner corpus analysis and the development of foreign language proficiency. *System, 32*, 577–591.

Carrió Pastor, M. L. (2004). *Contrastive analysis of scientific-technical discourse: Common writing errors and variations in the use of English as a non-native language.* Ann Arbour: UMI.

Carrió, M. L. & Seiz, R. (2000). La expresión escrita en inglés técnico. Sus errores. In F. Luttikhuizen (Coord.), *III Congreso Internacional sobre Lenguas para Finalidades Específicas* (pp. 69–73). Barcelona: Universidad de Barcelona.

Carrió, M. L. & Mestre, E. M. (2010). Implications of a corpus of lexical errors in second language learning. In I. Moskowich, B. Crespo, I. Lareo & P. Crespo (Eds.), *Proceedings of the II International Congress*. *AELINCO* (pp. 183‑196). A Coruña: Servicio de Publicaciones de la Universidad de A Coruña.

Corder, S. P. (1967). The significance of learner's errors. *International Review of Applied Linguistics, 5*, 161–170.

Chapelle, C. A. (2004). Technology and second language learning: Expanding methods and agendas. *System*, 32, 593–601.

Gaskell, D. & Cobb, T. (2004). Can learners use concordance feedback for writing errors? *System, 32*, 301–319.

Gotti, M. (2003). *Specialized Discourse. Linguistic Features and Changing Conventions*. Bern: Peter Lang.

Hamid, O. (2007). Identifying second language errors: how plausible are plausible reconstructions? *ELT Journal, 61*(2), 107–116.

Hunston, S. & Francis, G. (2000). *Pattern grammar: A corpus-driven approach to the lexical grammar of English.* Amsterdam: John Benjamins.

James, C. (1998). *Errors in language learning and use*. London: Longman.

Krishnamurthy, R. & Kosem, I. (2007). Issues in creating a corpus for EAP pedagogy and research. *Journal of English for Academic Purposes, 6*, 356–373.

Larsen-Freeman, D. & Long, M. H. (1992). *An introduction to second language acquisition research.* London: Longman.

Levison, M., Lessard, G. & Walker, D. (2000). A multi-level approach to the detection of second language learner errors. *Literary and Linguistic Computing*, *15*(3), 313–322.

Lee, I. (2004). Error correction in L2 secondary writing classrooms: The case of Hong Kong. *Journal of Second Language Writing, 13*, 285–312.

Mestre, E. M. (2011). *CEFR & Error analysis in second language teaching at university level.* Saarbrücken: Lambert Academic Publishing

Mudraya, O. (2006). Engineering English: A lexical frequency instructional model. *English for Specific Purposes, 25*, 235–256.

Nelson, M. (2006). Semantic associations in Business English: A corpus-based analysis. *English for Specific Purposes, 25*, 217–234.

Richards, J. (1971). A non-contrastive approach to error analysis. *English Language Teaching Journal, 25*, 204–219.

Richards, J. (1974). *Error Analysis*. London: Longman.

Rifkin, B. & Roberts, F. D. (1995). Error gravity: A critical review of research design (review article). *Language Learning*, 45(3), 511–537.

Salem, I. (2007). The lexical-grammatical continuum viewed through student error. *ELT Journal, 61*(3), 211–219.

Scott, M. (2009). *WordSmith Tools 5.0*. Liverpool: Lexical Analysis Software.

Webber, P. (1993). Writing medical articles: a discussion of common errors made by L2 authors and some particular features of discourse. *UNESCO-ALSED LSP Newsletter, 15*(2), 38–49.

Yoshii, M. (2006). L1 and L2 glosses: Their effects on incidental vocabulary learning. *Language Learning & Technology, 10*(3), 85–101.