Detecting Learning Disorders in Students' Written Production in the Foreign Language: Are Learner Corpora of Any Help?

MARÍA BELÉN DÍEZ BEDMAR
Departamento de Filología Inglesa
Universidad de Jaén

ABSTRACT: This paper reports on the possibility to detect learning disorders when conducting Computer-aided Error Analysis (CEA). The study of a longitudinal learner corpus compiled at the *Universidad de Jaén* (Spain) and error-tagged with the *Université Louvain Error Editor* (Hutchinson, 1996; Dagneaux, Denness, Granger and Meunier, 1996) reveals that the disorder of written expression, i.e. dysorthographia, may be detected in the students' production in the foreign language. Failure in recognising symptoms of this disorder may lead to incorrect interpretations of the CEAs, as the errors found in a learner corpus may be misleadingly attributed to the student's language acquisition process rather than to his or her learning disorders.
Keywords: disorders in written expression, dysorthographia, Computer-aided Error Analysis, longitudinal learner corpus.

Detección de trastornos del aprendizaje en la producción escrita de los alumnos en lengua extranjera: ¿Pueden ser útiles los corpus de estudiantes?

RESUMEN: Este artículo muestra la posibilidad de detectar trastornos de aprendizaje a la hora de realizar un análisis de errores informatizado. El estudio de un corpus de estudiantes longitudinal, compilado en la Univesidad de Jaén (España) y anotado con el anotador *Université Louvain Error Editor* (Hutchinson, 1996; Dagneaux, Denness, Granger and Meunier, 1996) muestra que es posible detectar un trastorno de la expresión escrita, disortografía, en la producción de estudiantes de inglés como lengua extranjera. Si los síntomas de este trastorno no se detectan, las interpretaciones de los resultados del análisis de errores informatizado pueden ser incorrectas, ya que los errores encontrados en un corpus de estudiantes se pueden atribuir de forma errónea al proceso de aprendizaje de idiomas del sujeto y no a su trastorno de aprendizaje.
Palabras clave: trastornos de la expresión escrita, disortografía, análisis de errores informatizado, corpus de estudiantes longitudinal

1. INTRODUCTION

(Computer) learner corpora (CLC), or learner corpora for short, have been used to describe the foreign or second language learners' written or oral production so that experts in second language acquisition (SLA) and foreign or second language teachers may obtain an in-

depth description of the students' process(es) of language acquisition, improve their curriculum, the teaching materials implemented, etc. The methodology frequently used has been Contrastive Interlanguage Analysis (CIA) (Granger, 1996: 43-44), Computer-aided Error Analysis (CEA) (Dagneaux, Denness, and Granger, 1998), or the Integrated Contrastive Method (ICM) (Granger, 1996: 46; Gilquin, 2000/2001) (see section 2).

If the analysis of the errors[1] made by a group of students is to be undertaken, the methodology which is frequently used for a CEA implies the hard task of error-tagging a (longitudinal) learner corpus (cf. Meunier, 1998: 19-37; Granger, 2002: 16-17; Prat Zagrebelski, 2004: 93; etc.), a time-consuming process in which the annotator detects, annotates and classifies the errors made by the learners and, therefore, becomes familiar with the most important type of errors which students have. After the error-tagging process, statistical analyses are run so that the statistically significant errors are obtained. With these findings, it is possible to determine the focus of future qualitative analyses and/or the aspects of the foreign language which will be dealt with in remedial teaching sessions or materials in a delayed human intervention (DHU) (Sinclair, 2001: xi) or, in other words, a synchronous corpus building project (Granger, 2004: 131).

However, this process normally misses a crucial aspect for the description of the students' interlanguage, since it does not normally distinguish between those errors which stem from their interlanguage level, and those errors which are caused because of the students' possible learning disorders. The main reason for this methodological drawback is that learning disorders are diagnosed and studied in the L1 (see section 4), if they are detected at all. In fact, it has long been claimed that pre- and in-service teachers lack experience in the recognition of students with special needs (cf. LeLoup and Ponterio, 1997: 1). Therefore, if it is difficult to detect «symptoms» of the students' learning disorders in the L1, the task proves almost impossible in their second or foreign language, the only possibility being the CEA analyses conducted on a learner corpus. It is for this reason that the role of the annotator of a learner corpus becomes crucial when there is no previous information on possible learning disorders, since the experience gained when error-tagging the learner corpus is decisive when determining which errors are frequent in the students' production as a group, and which ones are particular to a learner, even before conducting the statistical analyses.

However, the CEA analyses conducted so far have not proved the possibility to connect learner corpora and clinical linguistics, since the presence of any «symptoms» of learning disorders had not been reported until now, possibly due to two reasons: first, the level of granularity of the CEAs conducted have not considered in detail aspects which may be affected by learning disorders, such as spelling (see Section 3); and second, the learner corpora used lack an important number of words produced by the same learner, so that enough evidence of a student's possible learning disorder cannot be obtained and analysed - hence the importance of longitudinal learner corpora in this respect. As a consequence, participants who contribute their writings to learner corpora may have some type of learning disorder, and the errors made by these students may be misleadingly attributed to a language learning process, rather than a learning problem. This drawback may have two important consequences. The first one

1. Although there are differences between errors and mistakes (cf. Corder, 1967), the term error will be used throughout this paper.

affects the overall description of the student's interlanguage level, which may be wrong if the errors caused by learning disorders are not considered and distinguished from errors which stem from the student's interlanguage stage. The second consequence is closely related to the learner him/herself, because it is possible to contact him/her when «symptoms» appear in his/her writing so that s/he can seek appropriate help to be diagnosed and, if necessary, receive the appropriate treatment.

That was the case when error-tagging a four-year longitudinal learner corpus (283,623 words) compiled and error-tagged at the University of Jaén (Spain). When doing so, the annotator noticed that some of the spelling errors by a student were particular to this participant in the learner corpus, which was hypothesized to be due to a learning disorder. When asked about this possibility, two developmental psychologists pointed to a possible disorder of written expression (also known as dysorthographia), which is closely related to a reading disorder (i.e. dyslexia).[2] Despite the fact that dysorthographia and dyslexia normally appear together and one affects the other, the study of the student's possible learning disorders could not consider dyslexia because it was impossible to access the learner's reading skill at that time. Consequently, the study conducted only focuses on the disorder of written expression or dysorthographia, as defined by the Inserm Collective Expert Review (2007: 20):

> […] a specific disorder of spelling which accompanies dyslexia; the cognitive dysfunction underlying the two disorders is probably common to both. In dysorthographia, the spelling of words is highly deficient, a direct consequence of the phonological disorder in dyslexic children. (Inserm Collective Expert Review, 2007: 20).

Since the student was not aware of the possibility of having this learning disorder and had not been diagnosed by a psychologist at that time,[3] the main objectives of this paper were the following: first, to conduct a qualitative analysis of the spelling errors which were hypothesized to be caused by a learning disorder, with the help of two developmental psychologists; second, to compare the occurrence of those errors in the production of this particular student and in the production by her class-mates, i.e. the control corpus; and, finally, to classify the student's possible dysorthographic errors following Rivas Torres and Fernández Fernández's (2004) classification of dysorthographic subjects' writing and spelling, to see if those errors were really attributed to dysorthographia.

In order to explain how the hypothesis that this learner showed possible dysorthographic spelling errors[4] was confirmed, so that other researchers can replicate it with the spelling errors

---

2. For the classification of clinical association, as approved by the American Psychiatric Association, please refer to the DSM-IV-TR (American Psychiatric Association, 2000).

3. Since the researcher had this student's information in the learner profile form, it was possible to contact the student, who had already graduated, and inform her of such possibility in case she wanted to seek help from a psychologist. Some months later, the student was diagnosed with dyslexia and received the appropriate treatment.

4. The spelling errors of this student are referred to as 'possible dysorthographic spelling errors' for two reasons. First, the student was diagnosed after finishing her degree, that is, after providing all her samples for the longitudinal CLC. Therefore, it cannot be scientifically claimed that those errors are caused by dysorthographia

in their learner corpora in case spelling errors of this type are found, the methodology employed in the study is found in Section 5, and the results obtained are provided in sections 6 to 7, where the student' particular spelling errors are pointed out, compared to the control corpus and classified. Apart from the analysis of the possible dysorthographic spelling errors in this student's production, sections 2 to 4 familiarize the reader with basic information on computer learner corpora (section 2), CEA studies on spelling errors (section 3), and some publications on clinical linguistics, i.e. on the study of the language used by patients with some type of disorder (section 4). Finally, the paper draws some conclusions on the importance of being aware of the possibility to find some errors triggered by learning disorders when analysing the results of CEAs, so that results are not biased and the possibility to help students in their language learning process is considered.

2. ANALYSING FOREIGN OR SECOND LANGUAGE LEARNERS' PRODUCTION: CLC

The arrival of computer learner corpora (CLC), i.e. «[…] electronic collections of foreign or second language learner texts collected on the basis of strict design criteria» (Granger, Kraif, Ponton, Antoniadis and Zampa, 2007: 254), fostered research in the quantitative and qualitative analysis of second or foreign language learners' written or oral production, as can be seen in the ample number of studies published in specialised international journals and various edited books (cf. Granger, 1998; Granger, Hung and Petch-Tyson, 2002; Guy, Bernardini and Stewart, 2004; Sinclair, 2004; Gilquin, Papp and Díez-Bedmar, 2008; etc.).

Although some attempts had been made in the past to compile learner corpora, such as the project in *foreign language pedagogy corpus* (cited in Tono, 2002: 63-65 and Granger, 1993: 66), the *Zweitspracherwerb Italienischer, Spanischer und Portugiesischer Arbeiter* (*The ZISA project*) (Meisel, Clahsen and Pienemann, 1981), or *The European Science Foundation second language database* (*ESF database*) (Perdue, 1993), one of the most important research centres concerning the development of learner corpora nowadays is the *Centre for English Corpus Linguistics* at the *Université Catholique de Louvain* (Belgium), directed by Professor Granger. The compilation of the *International Corpus of Learner English* (*ICLE*), and its first release with the written production by learners from eleven different mother tongue backgrounds (Granger, Dagneaux and Meunier, 2002), proved to be an influential material which fostered the proliferation of further studies on the study of students' interlanguage. Following this first release, the second one, containing sixteen mother tongue backgrounds (Granger, Dagneaux, Meunier and Paquot, 2009), and the two new initiatives, namely the compilation of a learner corpus of English for Academic Purposes (the *VESPA* project) and a longitudinal learner corpus (the *LONGDALE* project),[5] provide the research field with the opportunity to collaborate and continue doing research in the production by students of English as a foreign language.

in case she developed it later. Second, the spelling errors described in this paper are found in the students' foreign language, whereas she was diagnosed in her L1.

5. For further information on these projects, please refer to the web page of the Centre for English Corpus Linguistics (http://cecl.fltr.ucl.ac.be/).

The methodologies employed for such studies normally involve the comparison of the production by a group of students with a control corpus or another group of students with another L1, i.e. a Contrastive Interlanguage Analysis (CIA) (Granger 1996: 43-44), or the use of a corpus of the students' L1 by means of the Integrated Contrastive Model (ICM) (Granger, 1996: 46; Gilquin 2000/2001: 100-101). These approaches to the study of learner language may also involve the analysis of the students' errors by means of a previous Computer-aided Error Analysis (CEA) (Dagneaux, Dennes and Granger 1998).[6] With this objective in mind, various error taxonomies have been developed (see, among others, Corder 1973: 275-285; Dulay, Burt and Krashen 1982: 146-197; James 1998: 102-113; 179-199; Nicholls, 2003; etc.), and have been used in numerous studies. However, the use of different error taxonomies has led to the unfeasible comparison of results, since the results of each CEA are highly determined by the error taxonomy applied, i.e. they depend on the classification of errors employed.

In order to compare the CEAs conducted on CLC, an attempt to standardize the error taxonomies used was also made by the *Centre of English Corpus Linguistics*. The result was the publication of the *Université Catholique de Louvain Error Editor* (*UCLEE*) (Hutchinson, 1996), and an accompanying Error Tagging Manual, where an error taxonomy and six error-tagging principles are provided (Dagneaux, Denness, Granger and Meunier, 1996), which has been used by researchers around the world (cf. Prat Zagrebelsky, 2004; Díez-Bedmar, 2005; Kaszubski, 2005; Ballesteros, Rica, Neff and Díez Prados, 2006; etc.). In this taxonomy, spelling errors (FS) are included within the broad category «Form», which also includes morphological errors. However, the studies which have provided some type of information on spelling by means of a CEA either with this error taxonomy or any other (see section 3) have not previously reported on symptoms of dysorthographic errors or the need to distinguish between those errors which are attributable to the language learning process and those which stem from a learning disorder.

3. The study of spelling in CLC

As previously claimed, spelling errors have not been widely studied in the production of students of English as a foreign language (cf. Nesselhauf, 2004: 134; 2006: 150). Four main reasons can be put forward for this lack of CLC-based studies on this aspect of the language. First, the use of electronic formats to hand in the students' written production may affect the number of spelling errors in the learner corpus because of students' possible use of spelling-checker programmes. Second, and as advocated in an early paper by Granger (1993: 61-62),[7] the normalization of some errors, among which spelling errors are found, was proved to

---

6. Although the focus of this paper is the detection of learning disorders by means of CEA, it is necessary to remember here that the analysis of students' learner production may not entail a CEA if students' errors alone are not under analysis. Thus, textual aspects related to the complexity, fluency or accuracy of the students' production may be used to describe the students' performance (cf. Wolfe-Quintero, Inagaki and Kim, 1998; Celaya Villanueva, Pérez Vidal and Torras Cherta, 2000-2001; Ortega, 2003, de Haan and Van Esch, 2008, etc.), or the holistic score of aspects related to the students' production (cf. Corpas and Madrid, 2007), may be used.

7. She refuted this claim later, so that the students' verbatim production was kept and analysed (Granger 1998: 17; Granger, Dagneaux and Meunier 2002: 20).

facilitate the description of the learners' production (cf. van Rooy and Schäfer, 2003), which led some researchers to do so (cf. Abe and Tono, 2005). Third, the (longitudinal) corpus-based analyses normally use oral data (cf. Lakshmanan and Selinker, 2001; Housen, 2000; 2002; Myles, 2005; Trenkic, 2007; etc.), thus resulting in the researcher's impossibility to establish whether the learner had spelling problems or not. Finally, the degree of granularity of the error taxonomy and the extent to which a learner corpus is error-tagged in full determine the inclusion of some aspects of the language, and spelling has frequently been neglected.

Whereas the first three reasons limit the number of studies on spelling due to their normalization or the use of oral learner data, the last one stems from the researcher's decision to include this aspect in the CEA. For example, three CEAs conducted on different sub-corpora of the *ICLE* (see section 2) only revealed the percentage of errors in the general error category «Form», which includes errors related to spelling and morphology.[8]

In the first one, a section of the Italian subcomponent of the *ICLE*, i.e. c. 50,000 words of the *ICLE-IT*, was analysed by Prat Zagrebelski (2004; 2005). As a result, 6.34% of the total amount of errors was claimed to correspond to the general category «Form», but no further mention is made to the part which corresponds to morphology or spelling errors. It is only in Prat Zagrebelski (2004) where the author mentions that «[…] there is a lot of carelessness in the use of capital letters and in the spelling of words […]» (Prat Zagrebelski, 2004: 93). Similarly, Kaszubski (2005) only reported the percentage of errors in the general error category «Form», which accounted for 8.90% of the total amount of errors in the Polish subcomponent of the *ICLE*, *PICLE*. Finally, Ballesteros, Rica, Neff and Díez Prados (2006) reported on 10% of formal errors in the *SPICLE*, i.e. the Spanish subcomponent of the *ICLE*.

However, two CEAs have provided the percentages of spelling errors in two learner corpora, instead of only providing the figures for the general error categories (cf. Díez-Bedmar, 2005 and Chuang and Nesi, 2006). The former analysed the written production by Spanish first-year university students using the *UCLEE* and the accompanying error manual, finding that spelling problems accounted for 6.12% of the errors in her learner corpus. The latter followed a taxonomy derived from Dulay, Burt and Krashen's (1982) classification of errors, James' (1998) addition of omission, overinclusion, misselection, missordering and blend, and the error categories in the *UCLEE* (see section 2), to find that misspelling (included in the lexical category in their error tagset) accounted for 13.4% of the total amount of errors made by Chinese students of English.

Apart from these papers, a remark is made in passim on the Spanish students' tendency to spelling errors in sequences such as «cu», instead of «qu», if compared to their Dutch, French and Polish counterparts in the *ICLEE*, and native speakers in the *LOCNESS* (Granger and Wynne, 1999).

As seen in this section, the number of CEA studies on spelling errors by students of English as a foreign language is limited, and the degree of granularity achieved in the error-

---

8. The seven general error categories in the error taxonomy designed by the Centre for Corpus Linguistics are the following: i) Form; ii) Grammar; iii) Lexico-Grammar; iv) Lexis; v) Word Redundant, Word Missing and Word order; vi) Register; and vii) Style. The general error category «Form» includes both «Morphology» and «Spelling» (cf. Dagneaux, Denness, Granger and Meunier 1996: 8-9).

tagging process is not enough to conduct qualitative analyses and find cases of possible dysorthographic errors. To the best of my knowledge, this is the first study which reveals learning disorders when conducting a CEA in a learner corpus.

4. Describing language disorders in the L1: clinical linguistics

Back in the 1990s, the language produced by second or foreign language learners and by patients with any disorder was considered as a «special language» (Sinclair, 1995: 24). As seen in sections 2 and 3 above, the study of the language produced by students has been the focus of interest of SLA researchers or teachers interested in improving their students' performance, but no studies had previously reported the presence of symptoms of any disorder in the production of students, i.e. the possible connection between learner corpora and clinical linguistics. However, the analysis of the language used by patients with any type of disorder is analysed by clinical linguists (cf. Crystal, 1992[2]; 2001), normally in close collaboration with neurophysiologists' and neurologists' to claim

> […] that, if in the abnormal brain we find differential impairments to different parts
> of language or components of the grammar, then such independent parts must exist
> in the normal intact brain and grammar; this leads to a greater understanding of the
> biological bases for these parts of the grammar (Fromkin, 1987: 5).

Thus, clinical linguistics has been used to analyse the language employed by subjects with various disorders, as seen in studies on aphasic language (Jakobson, 1955; 1968), Alzheimer patients (Almor, Kempler, MacDonald, Andersen and Tyler, 1999; Blair, Marczinksi, Davis-Faroque and Kertesz, 2007), autistic subjects (Capps, Kehres and Sigman, 1998; Solomon, 2004), schizophrenics (Bearden, Rosso, Hollister, Sanchez, Hadley and Cannon, 2000), Down's syndrome patients (Fowler, 1990; Ring and Clashen, 2005; among others), dyslexia (Fromkin, 1985; 1987; Pedler, 2007), etc.

In order to conduct such studies, data have been mainly obtained by elicitation, and also following a corpus linguistics methodology, as can be seen in the corpora compiled by Fletcher and Garman (1988) with normal and impaired language development, the *CHILDES* database (MacWhinney, 1996) with normal and impaired child language data in various L1s, or Perkins and Varley's (1996) *A Machine-Readable Corpus of Aphasic Discourse*. In Spain, the *Koiné* corpus (Otero Cabarcos and Fernández Casas, 2005; Souto Gómez, 2006), or the *PerLA* corpus (cf. Gallardo and Moreno, 2005; Moreno, 2008) have also been compiled to describe the language of aphasic people.

5. Methodology

For the purpose of this paper, the 4-year longitudinal learner corpus compiled at the University of Jaén (283,623 words) was divided into two parts. The first one consists of the handwritten production by one student of the English Studies degree (*Filología Inglesa*) who showed symptoms of dysorthographia in her production in the foreign

language.[9] Such section of the corpus amounts to 25,096 words, distributed along the four years as shown in Table 1:

Table 1. Diagnosed student's section.

Academic year	Number of words
1st	2,948
2nd	6,413
3rd	9,078
4th	6,656
	25,096

The control corpus contains the handwritten production by fifteen students who began the degree and proceeded along it during the same successive years as the diagnosed student. In this case, there are 258,527 words distributed as follows:

Table 2. Control computer learner corpus

Academic year	Number of words
1st	31,518
2nd	59,798
3rd	63,057
4th	104,154
	258,527

The complete learner corpus was compiled, transcribed and error-tagged in full with the Error Editor developed at the *Centre for English Corpus Linguistics* (Hutchinson, 1996) and its accompanying Error Tagging Manual (Dagneaux, Denness, Granger and Meunier, 1996).[10] Accordingly, spelling errors were annotated with the error tag «Form – Spelling» (FS) and the Concordance Tool in *WordSmith Tools* v.3 (Scott, 1998) was used to retrieve all the instances of misspelling.

A qualitative analysis followed to analyze if the spelling errors hypothesized to stem from the learning disorder were common to all the students in the learner corpus, i.e the sixteen students who contributed their production to the learner corpus or, on the contrary, they were peculiar to that particular student and showed the characteristics or symptoms of dysorthographia, as stated in Rivas Torres and Fernández Fernández (2004: 105-108).

9. This student was later diagnosed with dysorthographia based on DSM/IV/TR criteria (2000), i.e. the classification and description of clinical disorders by the American Psychiatric Association.
10. A native speaker of English highlighted the errors in the learner corpus.

6. Qualitative analysis: possible dysorthographic spelling errors

The first step to analyse the student's possible instances of dysorthographic spelling was to retrieve the concordance lines with her spelling errors. Then, each case of misspelling underwent a qualitative analysis to determine if it was an instance of a possible dysorthographic spelling with the assistance of two members of the Department of Psychology at the University of Jaén.

As a result, the thirty-seven types in the grey cells in Table 3 (with their correct spellings below for easier reference) were identified. When analysing them (see figures 1 and 2 for a couple of handwritten examples), it can be noticed that they are frequent words in any student's repertoire. Apart from words clearly related to the field of Linguistics, which should be mastered by any student taking the degree in English Studies, i.e. «declarative», «phonology» or «semantics», etc., most words do not belong to any specialised field.

Although on most occasions each type only occurs once, there are twelve types which present various tokens, as indicated in the number between brackets. Hence, *concieve*, *critize*, *fordward*, *posite* and *Romants* present two tokens, *ect* three, *believe* and *nowdays* four, *litle* and *though* five, and, finally, *masculine* presents six. Nevertheless, *poety* is the type presenting the highest number of tokens (22).[11]

Table 3. Possible dysorthographic spelling errors.

alway	*beth*	*belive* (4)	*charactertics*	*comparision*	*concieve* (2)	*converstio*
always	bed	believe	characteristics	comparison	conceive	conversion
critize (2)	*declarate*	*demosnstate*	*elicitist*	*ect* (3)	*exampe*	*expansionsim*
criticise	declarative	demonstrate	elitist	etc	example	expansionism
exploteted	*feates*	*flowe*	*fordward* (2)	*intest*	*intonatio*	*litle* (5)
exploded	features	flower	forward	interest	intonation	Little
masculin (6)	*mixting*	*nowdays* (4)	*peson*	*phology*	*pictres*	*poety* (22)
masculine	mixing	nowadays	person	phonology	pictures	poetry
posite (2)	*recived*	*romants* (2)	*secon*	*semantic*	*speakin*	*thougth* (5)
positive	received	romantics[12]	second	semantics	speaking	thought
undesrtant	*worllds*					
understand	words					

11. The misspelt words will be presented in italics for easier recognition
12. The use of *romantics* here refers to the poets in the romantic period.

Figure 1. poety.
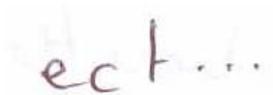
Figure 2. etc.

Once the words with possible dysorthographic misspelling were identified, a comparative study of the frequency of the thirty-seven types in Table 3 when correctly spelt and when showing the spelling above was conducted in the student's section of the computer learner corpus. In this way, it would be possible to investigate if they were cases of slips of the pen rather than errors attributable to dysorthographia.

As seen in Figure 3, the instances of possible dysorthographic spelling of fourteen words outnumber the number of correctly spelt instances. In fact, if percentages are considered, the words which are always incorrectly spelt, possibly due to dysorthographia, are *comparision*, *concieve*, *critize*, *exploteted*, *fordward*, *masculin*, *mixting*, *pictres* and *recieve*, followed by *ect* (60%), *converstio*, *demosnstate*, *nowdays*, and *phology* (50% each).
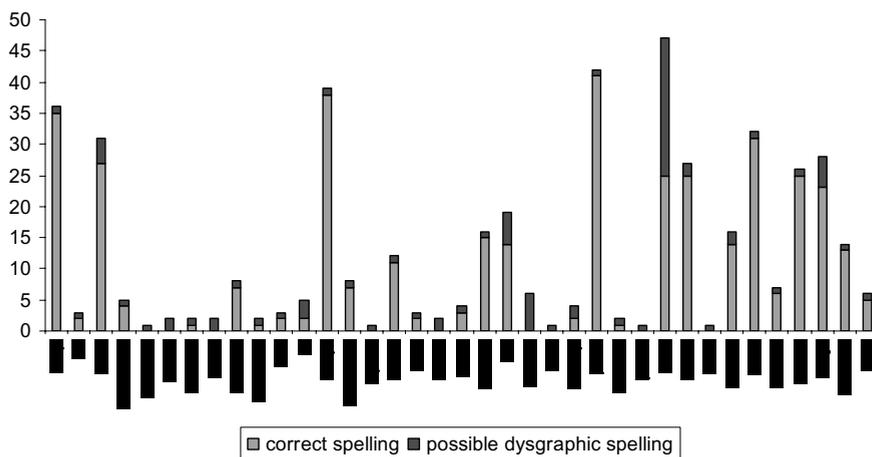

Figure 3. Number of cases presenting possible dysorthographic spellings per word.

These results may point to the relevance of those fourteen words, specially the ones which were always misspelt, i.e. *comparision*, *concieve*, *critize*, *exploteted*, *fordward*, *masculin*, *mixting*, *pictres* and *recieve*, when considering dysorthographic problems. However, the other twenty-three words which were recognised as possibly having spelling problems because of dysorthographia should also be taken into account, as they can follow a specific misspelling pattern (see section 8).

44

7. DETERMINING POSSIBLE DYSORTHOGRAPHIC SPELLING

In order to determine if the possible dysorthographic spelling errors in the student's written production were particular to this student or common to university students of English Studies at the University of Jaén, concordances with the thirty-seven types in Table 3 above were run in the control corpus.
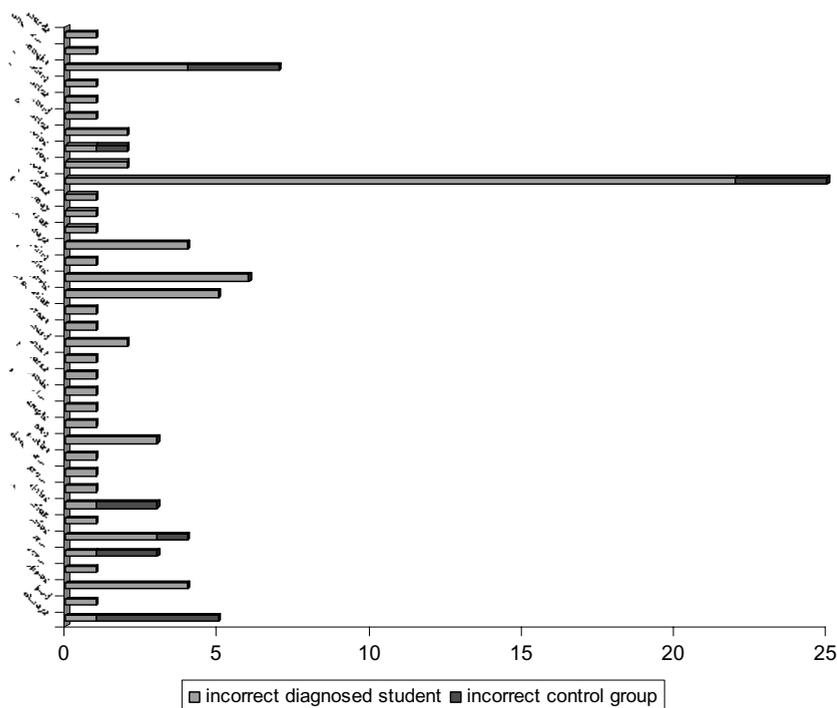
Figure 4. Possible dysorthographic spelling errors by the student and the control group.

Figure 4 shows that most cases of spelling errors are particular to that student, whereas the spelling errors in seven types, namely *alway*, *comparision*, *concieve*, *critize*, *poety*, *recieve* and *thougth*, can also be seen in the written production by the control group. Nevertheless, it is important to remark that the number of possible dysorthographic spelling errors in the control corpus is sixteen (in the production of the fifteen students along the four-year degree), which is significantly inferior to the eighty-four tokens in the student's production.

In the light of these results, it may be claimed that the eighty-four possible spelling errors (within thirty-seven types) in the written production by that particular student are symptomatic of her learning disorder, despite the fact that sixteen of them (in seven types) are also found in the 258,527 words in the control corpus.

8. Classification of the possible dysorthographic spelling errors

The thirty-seven possible dysorthographic spelling problems found in the student's written production were further examined to see if there was any repeated misspelling pattern. In order to do so, Rivas Torres and Fernández Fernández's (2004: 105-108) classification of the systematic and repetitive errors of the dysorthographic subjects' writing and spelling was considered. This taxonomy identifies errors regarding their linguistic-perceptive, visuospatial or visuoauditory origin, their content and spelling rules. Among them, especially relevant to this student's written production was the first one, i.e. the errors with a linguistic-perceptive origin, which is further subdivided into errors of omission and addition of graphemes or syllables, as well as the substitution and reversal of graphemes, among others.

In fact, the analysis of the forty-one spelling errors in the thirty-seven types found in the written production by this student (cf. Table 3) could be classified into the linguistic-perceptive category,[13] which seems to point out to the dysorthographic origin of these errors. As revealed in Figure 5 below, 61% of the total amount of possible dysorthographic spelling errors is related to omission (25 instances), then followed by addition (7), reversal (5), and, finally, substitution (4).
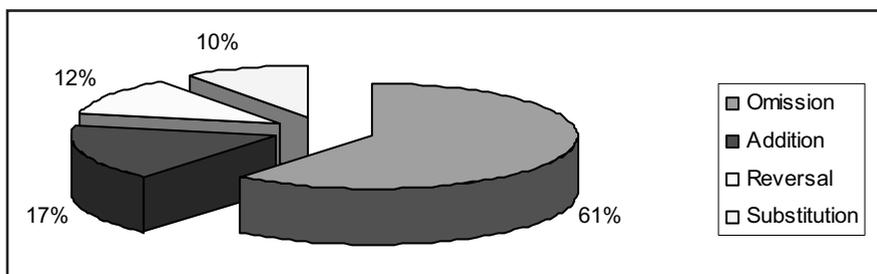

Figure 5. Breakdown of the student's linguistic-perceptive errors.

Within the most frequent subcategory, the omission of single graphemes is the one which presents more instances: 10 in medial position and 6 at end position (cf. Figure 6). Therefore, the omission of more than one grapheme is half as frequent as the omission of single graphemes. When this is the case, the number of omitted graphemes is normally two, which may be together in the word (i.e. *phology*, *Romants*, *intest* and *criticing*), or in various positions within the word (i.e. *sematic* and *converstio*). Finally, there are two cases in which more than two graphemes are omitted, as can be seen in *posite* and *declarate*.

Whereas the previous twenty-four omissions involve graphemes, the omission of the syllable «is» in the word «characteristics» (char·ac·ter·is·tics) proves the only case of syllable omission. However, it is important to remark that the students' L1 and FL have a different

---

13. Even though there were thirty-seven types of spelling errors which may stem from dysorthographia (cf. Table 3), we consider here forty-one instances because *criticize* presents one dysorthographic problem in each of its two occurrences and *demonstrate, explode* and *understand* present two dysorthographic problems each.

syllable division: Spanish is a syllable-timed language, whereas English is stress-timed, which is related to the perceived phonetic similarity/distance between languages (cf. Alonso Marks, Bond and Stockmal, 2003: 29). Consequently, Spanish learners of English have difficulties when understanding and applying the concept of syllable in the foreign language. Then, the omission of two letters which correspond to different syllables is frequent (i.e. *phology*), the omission of the two letters which form a syllable being a random case (i.e. *charactertics*).
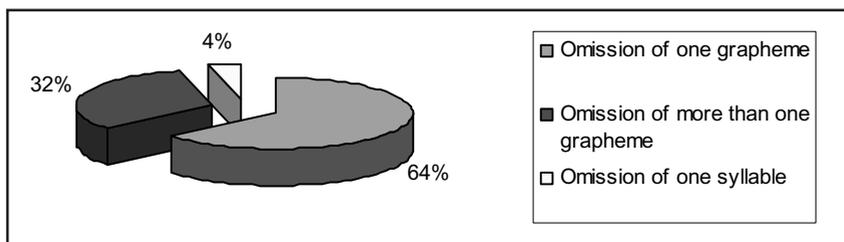
Figure 6. Omission: breakdown of possible dysorthographic spelling errors.

	Medial position	End position
Omission of single letters	*belive, demosnstate, exampe, feates, litle, nowdays, peson, pictres, poety, recived*	*alway, flowe, intonation, masculin, secon* (hand), *speakin*
	10	6
		16 (64%)
Omission of more than one letter	*criticing, declarate* (questions), *intest, phology, posite* (oriented), *Romants*	
	converstio, sematic,	
		8 (32%)
Omission of one syllable	*charactertics*	
		1 (4%)
Total omission		**25**

   Three examples stand out when analysing addition (see Figure 7 and Table 5). The first one is in the word *elicitist*, which presents the addition of «ci», which is not a syllable, but might be considered so by a Spanish learner of English. The other two are the only cases of repetition of one letter, i.e. *worllds*, or two, i.e. *exploteted*. As seen in the table below, all the cases of addition occur in medial position, and there is not any case of the addition of a syllable.
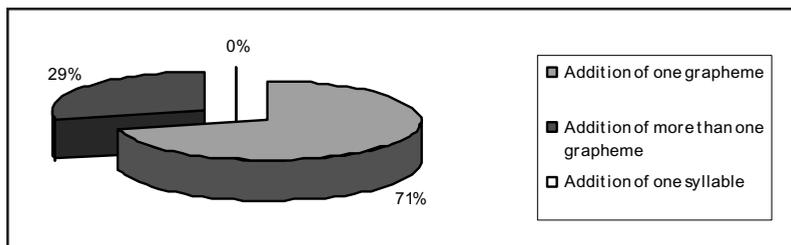
Figure 7. Addition: breakdown of possible dysorthographic spelling errors.

Table 5. Types of addition errors in the student's production

	Medial position	End position
Omission of single letters	*belive, demosnstate, exampe, feates, litle, nowdays, peson, pictres, poety, recived*	*alway, flowe, intonation, masculin, secon* (hand), *speakin*
	10	6
		16 (64%)
Omission of more than one letter	*criticing, declarate* (questions), *intest, phology, posite* (oriented), *Romants*	
	converstio, sematic,	
		8 (32%)
Omission of one syllable	*charactertics*	
		1 (4%)
Total omission		**25**

Finally, reversal and substitution of phonemes can be noticed in the student's written production in medial and end position (Tables 6 and 7, respectively). In the cases of substitution, it can be noticed that related phonemes are prone to such substitution, as can be seen in the four cases: *exploteted* /t/ and /d/; *undesrtant*, /t/ and /d/; *critisize*, /s/ and /z/; and *beth*, /th/ and /d/.

Table 6. Types of reversal of phonemes in the student's production

	Medial position	End position
Reversal of phonemes	*concieve, expansionsim*	*thougth, undesrtant, ect*
Total reversal **5**		

Table 7. Types of substitution of phonemes in the student's production

	Medial position	End position
Substitution of phonemes	*exploteted*, *critize*	*beth*, *undesrtant*
Total substitution **4**		

9. Conclusions

Many publications have already reported on the use of CEA to analyse the students' interlanguage cross-sectionally or longitudinally to improve the description of the students' use of the language, which may inform learner-corpus-based teaching materials. However, apart from relevant issues such as the error taxonomy used, its granularity or the computer learner corpus used, there is an aspect which proves crucial to conduct a good CEA: a complete learners' profile form, i.e. one with as much information as possible, including the learner's learning disorders, if any. However, this piece of information is frequently missed, since those disorders are normally underdetected.

The omission of such information has twofold consequences, one regarding the students' career and the other concerning researchers' CEA results. As far as the former is concerned, students who are not aware of their learning needs may finish their university studies without being conscious of such reality, thus having more difficulties when facing their postgraduate studies, looking for a job or planning to take the official State exam to become secondary school teachers in Spain. As for the latter consequence, failure to recognize symptoms of a student's learning disorder may bias the interpretation of the CEAs conducted, since all the errors found in a learner corpus may be misleadingly attributed to the language acquisition process, thus affecting the quantitative and qualitative analysis of the student's interlanguage stage. In fact, that would have been the case with the eighty-four spelling errors of the particular student in this paper (see sections 6 to 8), if their repeated pattern had been overlooked by the error-annotator. As it may happen if learning disorders are not considered, these errors would have been added to the other ones which this student also makes (which are similar to the control corpus and typical of this interlanguage stage) and explained because of her particular acquisition process, rather than her (until then) underdetected learning disorder (dysorthographia). It is for this reason that longitudinal CLC, in which the production by the same student over several years is studied, prove to be a better source of data than cross-sectional ones: if the production by only one student in a year is considered, repeated patterns may not emerge and, therefore, learning disorders may go unnoticed.

This paper has proved how the initial annotator's hypothesis that this student showed dysorthographic spelling errors was confirmed by analysing the spelling errors that she had made from a qualitative point of view, and comparing them to those by her classmates. Together with this guide to help other researchers notice symptoms of the disorder of the written expression, the conclusions of this paper can be taken as a note of caution directed to annotators or researchers working with learner corpora so that i) information on the students' learning disorders, if any, is required in learner profile forms; ii) errors which stem from

learning disorders are not misleadingly attributed to the student's learning process but to his/her learning disorder; and iii) those students who have contributed to the learner corpus and show a possible learning disorder may seek psychological help and undergo the appropriate treatment. As can be seen, the role played by the annotator is crucial and proves a possible first step towards detection and diagnosis of students' learning disorders and, thus, a new approach to clinical linguistics by means of learner corpora.

Even though dysorthographia and dyslexia are closely related and normally diagnosed together, only dysorthographia, as probably revealed by learners' spelling errors, can be dealt with when using written CLC-based data. In order to conduct research on dyslexia, it is also necessary to access the students' reading skill.

Hence, researchers involved in error-tagging a (longitudinal) learner corpus need to be aware of the possibility of detecting some learning disorders: noticing errors which are particular to one student and follow a repeated pattern may arise the error annotator's suspicion of possible learning disorders, which can be later discussed in an interdisciplinary team and, if necessary, reported to the student. Thus, if the learner is part of an Early Human Intervention (EHI) study (Sinclair, 2001: xi), or in Granger's words, a synchronous corpus building project (Granger, 2004: 131), the participants who provide data for the research may benefit even more from the time spent and effort made in compiling, error-tagging and analysing that (longitudinal) learner corpus.

10. ACKNOWLEDGMENTS

11. REFERENCES

Abe, M. and Tono, Y. (2005). «Variations in L2 spoken and written English: Investigating patterns of gramatical errors across proficiency levels», in P. Danielsson and M. Wagenmakers (eds.), *Proceedings from the Corpus Linguistics Series* 1 (1), available from: http://www.corpus.bham.ac.uk/PCLC/, accessed 23 August, 2009.

Alonso Marks, E., Bond, Z. E. and Stockmal, V. (2003). «Language experience and the representation of phonology in an unknown language», in *Revista Española de Lingüística Aplicada*, *16*: 23-31.

Almor, A., Kempler, D., MacDonald, M. C., Andersen, E. S. and Tyler, L. K. (1999). «Why do Alzheimer patients have difficulty with pronouns? Working memory, semantics, and reference in comprehension and production in Alzheimer's disease», in *Brain and Language*, *67*, 3: 202-227.

American Psychiatric Association, APA (2004). *Diagnostic and Statistical Manual of Mental Disorders, Text Revision*. Washington, DC: APA.

Ballesteros, F., Rica, J.P., Neff, JA., and Díez Prados, M. (2006). «The ICLE error tagging project: analysis of Spanish EFL writers», in C. Mourón Figueroa and T. I. Moralejo Gárate (eds.), *Studies in Contrastive Linguistics. Proceedings of the 4th International Contrastive Linguistics Conference, September 2005*. Santiago de Compostela: Servizo de Publicacións e Intercambio Científico, 89-97.

Bearden, C.E., Rosso, I.M., Hollister, J.M., Sanchez, L.E., Hadley, T. and Cannon, T.D. (2000). «A prospective cohort study of childhood behavioural deviance and language abnormalities as predictors of adult schizophrenia», in *Schizophrenia Bulletin*, *26*, 2: 395-410.

Blair, M., Marczinski, C.A., Davis-Faroque, N. and Kertesz, A. (2007). «A longitudinal study of language decline in Alzheimer's disease and frontotemporal dementia», in *Journal of the International Neuropsychological Society*, *13*: 237-245.

Capps, L., Kehres, J. and Sigman, M. (1998). «Conversational abilities among children with autism and children with developmental delays», in *Autism*, *2*, 4: 325-244.

Celaya Villanueva, M.L., Pérez-Vidal, C. and Torras Cherta, M.R. (2000-2001). «Matriz de criterios de medición para la determinación del perfil de competencia lingüística escrita en inglés (LE)», in *RESLA*, 14: 87-98.

Chuang, F-Y. and Nesi, H. (2006). «An analysis of formal errors in a corpus of L2 English produced by Chinese students», in *Corpora*, *1*, 2: 251-271.

Corder, S.P. (1967). «The significance of learner's errors», in International Review of *Applied Linguistics*, *5*: 161-169.

Corder, P. (1973). *Introducing Applied Linguistics*. Great Britain: Penguin Education.

Corpas, M.D. and Madrid, D. (2007). «Desarrollo de la producción escrita en inglés al término de la Educación Secundaria Obligatoria española», in *Porta Linguarum*, *8*: 169-188.

Crystal, D. (1992). *Profiling Linguistic Disability*. London: Whurr Publishers.

Crystal, D. (2001). «Clinical Linguistics», in M. Aronoff and J. Rees-Miller (eds.), *The Handbook of Linguistics*. Oxford: Oxford University Press, 673-682.

Dagneaux, E., Denness, S. Granger, S. and Meunier, F. (1996). *Error Tagging Manual Version 1.1*. Centre for English Corpus Linguistics, Université Catholique de Louvain: Louvain la Neuve.

Dagneaux, E., Denness, S. and Granger, S. (1998). «Computer-aided error analysis», in *System*, *26*: 163-174.

de Haan, P. and Van Esch, K. (2008). «Measuring and assessing the development of foreign language writing competence», in *Porta Linguarum*, *9*: 7-21.

Díez-Bedmar, M.B. (2005). «Struggling with English at university level: error-patterns and problematic areas of first-year students' interlanguage», in P. Danielsson and M. Wagenmakers (eds.), *The Corpus Linguistics Conference Series*. [Internet document available at http://www.corpus.bham.ac.uk/PCLC/, last access 16/08/09].

Dulay, H., Burt, M. and Krashen, S. (1982). *Language Two*. Oxford: Oxford University Press.

Fletcher, P. and Garman, M. (1988). «Normal development and language impairment: syntax and beyond», in *Clinical Linguistics and Phonetics*, *2*: 97–113.

Fowler, A.E. (1990). «Language abilities in children with Down syndrome: evidence from a specific syntactic delay», in D. Cicchetti and M. Beeghly (eds.), *Children with Down Syndrome. A Developmental Perspective*. New York: Cambridge University Press, 302-228.

Fromkin, V. (1985). «Evidence in Linguistics», in S-Y. Killingley (ed.), *Linguistics and Linguistic Evidence*. Newcastle upon Tyne: Grevatt and Grevatt, 18-38.

Fromkin, V. (1987). «The lexicon: evidence from acquired dyslexia», in *Language*, *63*, 1: 1-22.

Gallardo, B. and Moreno, V. (2005). *Afasia no fluente*. Valencia: Guada Impresores.

Gilquin, G. (2000/2001). «The Integrated Contrastive Model. Spicing up your data», in *Languages in Contrast*, *3*, 1: 95-123.

Gilquin, G., Papp, Sz. and Díez-Bedmar, M.B. (eds.) (2008). *Linking up Contrastive and Learner Corpus Research*. Amsterdam and New York: Rodopi.

Granger, S. (1993). «International corpus of learner English», in J. Aarts, P. de Haan and N. Oostdijk (eds.), *Papers from the Thirteenth International Conference on English Language Research on Computerized Corpora, Nijmegen, 1992*. Amsterdam/Atlanta: Rodopi, 57-96.

Granger, S. (1996). «From CA to CIA and back: an integrated approach to computerized bilingual and learner corpora», in K. Aijmer, B. Altenberg and M. Johansson (eds.), *Languages in Contrast. Text-Based Cross-Linguistic Studies*. Lund: Lund University Press, 37-51.

Granger, S. (ed.) (1998). *Learner English on Computer*. London/New York: Addison Wesley Longman.

Granger, S. (2002). «A bird's-eye view of learner corpus research», in S. Granger, J. Hung and S. Petch-Tyson (eds.), *Computer Learner Corpora, Second Language Acquisition and Foreign Language Teaching*. Amsterdam/Philadelphia: John Benjamins, 3-36.

Granger, S. (2004). «Computer Corpus Research: Current Status and Future Prospects», in U. Connor and T. A. Upton (eds.), *Applied Corpus Linguistics. A Multidimensional Perspective*. Amsterdam/New York: Rodopi, 123-145.

Granger, S. and Wynne, M. (1999). «Optimising measures of lexical variation in EFL learner corpora», in J. Kirk (ed.), *Corpora Galore. Analyses and Techniques in Describing English. Papers from the Nineteenth International Conference on English Language Research on Computerized Corpora*. Amsterdam/Atlanta: Rodopi, 249-257.

Granger S., Dagneaux, E. and Meunier, F. (2002). *The International Corpus of Learner English*. Handbook and CD-ROM. Louvain-la-Neuve: Presses Universitaires de Louvain.

Granger, S., Hung, J. and Petch-Tyson, S. (eds.) (2002). *Computer Learner Corpora, Second Language Acquisition and Foreign Language Teaching*. Amsterdam/Philadelphia: John Benjamins.

Granger, S., Kraif, O., Ponton, C., Antoniadis, G. and Zampa, V. (2007). «Integrating learner corpora and natural language processing: A crucial step towards reconciling technological sophistication and pedagogical effectiveness», in *ReCALL*, *19*, 3: 252-268.

Granger, S., Dagneaux, E., Meunier, F. and Paquot, M. (2009). *International Corpus of Learner English, v. 2*. Louvain-la-Neuve: Presses Universitaires de Louvain.

Guy, A., Bernardini, S. and Stewart, D. (eds.). (2004). *Corpora and Language Learners*. Amsterdam/Philadelphia: John Benjamins.

Housen, A. (2000). «Verb semantics and the acquisition of tense-aspect in L2 English», in *Studia Linguistica*, *54*, 2: 249-259.

Housen, A. (2002). «A corpus-based study of the L2-acquisition of the English verb system», in S. Granger, J. Hung and S. Petch-Tyson (eds.), *Computer Learner Corpora, Second Language Acquisition and Foreign Language Teaching*. Amsterdam/Philadelphia: John Benjamins, 77-116.

Hutchinson, J. (1996). *UCL Error Editor*. Louvain-la-Neuve: Centre for English Corpus Linguistics, Université Catholique de Louvain.

Inserm Collective Expert Review (2007). «Dyslexia, Dysorthography, Dyscalculia. Review of Scientific Data», available from: http://ist.inserm.fr/basisrapports/dyslexie/dyslexie-synthese-anglais.pdf, accessed 24 August, 2009.

Jakobson, R. (1955). «Aphasia as a linguistic problem», in H. Werner (ed.), *On Expressive Language*. Worcester, MA: Clark University Press, 69-81.

Jakobson, R. (1968). *Child Language, Aphasia and Phonological Universals*. The Hague: Mouton.

James, C. (1998). *Errors in Language Learning and Use. Exploring Error Analysis*. London/ New York: Longman.

Kaszubski, P. (2005). «Typical errors of Polish advanced EFL learner writers», available from: http://ifa.amu.edu.pl/~kprzemek/concord2adv/errors/errors.htm, accessed 14 August, 2009.

Lakshamanan, U. and Selinker, L. (2001). «Analysing interlanguage: how do we know what learners know?», in *Second Language Research*, *17*, 4: 393-420.

LeLoup, J.W. and Ponterio, R. (1997). «On the net. Language Education and Learning Disabilities», in *Language Learning & Technology*, *1*, 1: 2-4.

MacWhinney, B. (1996). «The CHILDES system», in *American Journey of Speech-Language Pathology*, *5*: 5-14.

Meisel, J.M., Clahsen, H. and Pienemann, M. (1981). «On determining developmental stages in natural second language acquisition», Studies in *Second Language Acquisition*, *3*, 2: 109-35.

Meunier, F. (1998). «Computer tools for the analysis of learner corpora», in S. Granger (ed.), *Learner English on Computer*. London/ New York: Addison Wesley Longman, 19-37.

Myles, F. (2005). «Interlanguage corpora and second language acquisition research», in *Second Language Research*, *21*, 4: 373-391.

Moreno, V. (2008). «Application of PerLa aphasic corpus for the Speech Therapist teaching». Paper presented at New Trends in Language Teaching and Translation Studies: In Honour of John Sinclair, held at the University of Granada, 22-24 September 2008. Unpublished manuscript.

Nesselhauf, N. (2004). «How learner corpus analysis can contribute to language teaching: A study of support verb constructions», in G. Aston, S. Bernardini and D. Stewart (eds.), *Corpora and Language Learners*. Amsterdam/Philadelphia: John Benjamins, 109-124.

Nicholls, D. (2003). «The Cambridge learner corpus –error coding and analysis for lexicography and ELT», in D. Archer, P. Rayson, A. Wilson and T. McEnery (eds.), *Proceedings of the Corpus Linguistics 2003 Conference (28-31 March)*. Lancaster: University Centre for Computer Corpus Research on Language, 572-581.

Ortega, L. (2003). «Syntactic complexity measures and their relationship to L2 proficiency: a research synthesis of college-level L2 writing», in *Applied Linguistics*, *24*, 4: 492-518.

Otero Cabarcos, P. and Fernández Casas, M.X. (2005). «Algunos procesos de combinatoria sobre el análisis de corpus del habla infantil del grupo Koiné», in M. Veyrat Rigat and E. Serra Alegre (eds.), *Estudios de Lingüística Clínica (vol. 4). Problemas de eficacia comunicativa. Descripción, detección, rehabilitación*. València: Universitat de València/AVaLCC, 125-138.

Pedler, J. (2007). Computer correction of real-word spelling errors in dyslexic text. Unpublished PhD thesis. Birkbeck, University of London.

Perdue, C. (ed.) (1993). *Second Language Acquisition by Adult Immigrants. (2 vol)*. Cambridge: Cambridge University Press.

Perkins, M.R. and Varley, R. (1996). *A Machine-Readable Corpus of Aphasic Discourse*. University of Sheffield: Department of Human Communication Sciences/Institute for Language, Speech and Hearing (ILASH).

Prat Zagrebelsky, M.T. (2004). *Computer Learner Corpora. Theoretical Issues and Empirical Case Studies of Italian Advanced EFL Learners' Interlanguage*. Alessandria: Edizioni dell'Orso.

Prat Zagrebelsky, M.T. (2005). «Il computer learner corpora: una risorsa per la costruzione di attività di consapevolezza e di recupero per apprendetti di inglese come lengua straniera», in *Proceedings of the NAPLES Associazione Italiana Centri Linguistici Universitari (AICLU) Conference*, available from: http://www.cla.unina.it/ITALIANO/convegno/programma.htm, accessed 24 July, 2009.

Ring, M. and Clashen, H. (2005). «Distinct patterns of language impairment in Down's syndrome and Williams syndrome: the case of syntactic chains», in *Journal of Neurolinguistics*, *18*, 6: 479-501.

Rivas Torres, R.M. and Fernández Fernández, P. (2004). *Dislexia, disortografía y disgrafía*. Madrid: Pirámide.

Scott, M. (1998). *Wordsmith Tools Version 3*. Oxford: Oxford University Press.

Sinclair, J. (1995). «Corpus typology –a framework for classification», G. Melchers and B. Warren (eds.), *Studies in Anglistics*. Stockholm: Almqvist and Wiksell, 17-33.

Sinclair, J. (2001). «Preface», in M. Ghadessy, A. Henry and R. L. Roseberry (eds.), *Small Corpus Studies and ELT. Theory and Practice*. Amsterdam/Philadelphia: John Benjamins, vii-xv.

Sinclair, J. (ed.) (2004). *How to Use Corpora in Language Teaching*. Amsterdam/Philadelphia: John Benjamins.

Solomon, O. (2004). «Narrative introductions: discourse competence of children with autistic spectrum disorders», in *Discourse Studies*, *6*, 2: 253-276.

Souto Gómez, M. (2006). «El enfoque tipológico en afasiología: variación intelingüística de los trastornos afásicos», in *Actas del VII Congreso de Lingüística General (Barcelona, 18-21 abril de 2006)*. Barcelona: Universitat de Barcelona.

Tono, Y. (2002). The role of learner corpora in SLA research and foreign language teaching: the multiple comparison approach. Unpublished PhD thesis. Lancaster University.

Trenkic, D. (2007). «Variability in second language article production: beyond the representational deficit vs processing constraints debate», in *Second Language Research*, *23*, 3: 287-327.

Van Rooy, B. and Shäfer, L. (2003). «An evaluation of three POS taggers for the tagging of the Tswana Learner English corpus», in D. Archer, P. Rayson, A. Wilson and T. McEnery (eds.), *Proceedings of the Corpus Linguistics 2003 Conference (28-31 March)*. Lancaster University: University Centre for Computer Corpus Research on Language, 835-844.

Wolfe-Quintero, H., Inagaki, S. and H-Y. Kim (1998). *Second Language Development in Writing: Measures of Fluency, Accuracy and Complexity*. Honolulu: University of Hawaii Press.