Telling ELT Tales out of School

A learner corpus-based study on error associations

Ana Díaz-Negrillo*, Salvador Valeraa

* University of Jaén, Paraje Las Lagunillas s/n, Jaén, 23071, Spain

Abstract

This is a corpus-based study on error associations in English. Statistical analysis has proved powerful to reveal associations between patterns found in a multi-layered error-annotated English learner corpus by Spanish university students.

© 2010 Published by Elsevier Ltd. Open access under CC BY-NC-ND license.

Key words: ELT; learner corpora; error-tagging; error analysis; statistical analysis

1 This paper has been funded by the Spanish Ministry of Education (research project HUM2007-60107/FILO).

* Corresponding author. Tel.: +34-953-212135; fax: +34-953-211824. E-mail address: adinegri@ujaen.es
1. Introduction

The use of native and learner corpora has received increasing attention in English Language Teaching (ELT), as valuable resources in several ways (see Keck, 2004 for a review of the various applications of learner corpora in ELT or Campoy-Cubillo et al., 2010). One of the areas to explore with corpora and, specifically, with learner corpora is learner errors. Although it may have been under some criticism in SLA research, the study of learners’ errors remains of interest in Foreign Language Teaching (FLT) as it provides data on learners’ difficulties. Accordingly, attempts have been made to error-tag learner corpora with a view to conducting systematic explorations into learners’ errors (Díaz Negrillo, 2009; Hutchinson, 1996). Error analysis in learner corpora is, thus, an active research area (De Felice & Pullman, 2009; Gamon et al., 2009; Mendikoetxea et al., 2004), and some of the outcomes of error annotating corpora and the study of errors are, e.g. series like Common errors [...] and how to avoid them by Cambridge University Press, and learner corpora-informed dictionaries like the Longman active study dictionary (Manning, 2010). However, to our knowledge, there is little research on associations between errors in learner corpora.

This paper applies statistical analysis on a sample of the NON-native Corpus of English (NOCE, Díaz Negrillo, 2009) for any such associations. The NOCE corpus is a corpus of over 300,000 words of written English by Spanish undergraduates, contained in 1,054 samples of an average of 200 words each. The texts were collected from 2003 to 2009 primarily among first year students doing the English degree programme at the Universities of Granada and Jaén (Spain). The participants’ age is 18/19, and their average level of English is intermediate. The corpus samples were collected at three stages in the academic year (beginning, mid-term and end) to allow for longitudinal studies within each academic year. The samples were collected by the students’ lecturers assisted occasionally by the corpus compilers during 1-hour teaching sessions, always on a voluntary basis and under appropriate anonymity conditions. The samples were presented as a timed classroom task: it involved writing an essay on one of three topics suggested, all different for each year sampling, or on a fourth option of free writing for students who may not have enough background knowledge of the topics suggested, or may prefer not to discuss them. The corpus contains editorial and error annotation. Regarding the former, the entire corpus is annotated with the tagset EYES (Explicitly Encoded Surface modifications) for students’ edition of their own writing (struckouts, late insertions, etc.). A 40,000 word section of the corpus is error-tagged with EARS (Error-Annotation and Retrieval System, Díaz Negrillo, 2009), a corpus-driven tagset that classifies learner errors at 6 levels: i) Punctuation (PN), ii) Spelling (SP), iii) Word grammar (WG), iv) Phrase grammar (PG), v) Clause grammar (CG), and vi) Lexis (LX). Errors were annotated by 4 researchers and supervised by a native expert in learner corpus research. References on the 6 levels of description were used for annotation discrepancies, as described in Díaz Negrillo (2009). Examples of error tags are given below and after Tables 1 to 6 (showing only the tags relevant to the point at issue:

(1) [...] there are &WG.AD.NB.LU.I.T.MS>differents</WG.AD.NB.LU.I.T.MS> people [...] 
(2) He is a &WG.NN.NB.LU.ER.MS>students</WG.NN.NB.LU.ER.MS> of relation of laboral.
The tags include information about errors with respect to the following aspects:

1. Level of description: PN, SP, WG, PG, CG and LX; in (1) and (2) WG stands for Word grammar.
2. Type of unit: part of the speech, punctuation marks, grapheme type, etc. The parts of speech considered are adjective (AD), article (AR), adverb (AV), auxiliary verb (AX), conjunction (CJ), noun (NN), preposition (PE), pronoun (PO) and verb (VR). Thus, in (1) AD stands for Adjective and in (2) NN stands for Noun.
3. Linguistic category, function or aspect affected in the error: Upper case, Derivation, Tense, Foreign lexis, etc.; in (1) and (2) NB stands for Number and LU stands for Plural.
4. Nature of the error: execution (internal error, IT) or use of a function or category (external error, ER); in (1) IT is for Internal (adjectives do not take number); in (2) ER is for External, (plural occurs in an inappropriate context).
5. Surface modification: Misselection (MS), Omission (OM), Ordering (OR) and Overinclusion (OV); in (1) and (2) MS stands for Misselection.

2. Method

This paper uses a corpus section of samples by 30 students selected from the first and third batches of one of the academic years of the corpus study (2003/04). The corpus section thus contains 60 samples, approximately 12,500 words and 408 error tags. The corpus data were analyzed using two statistical methods: the chi-square test of independence and the Kendall rank correlation coefficient. The chi-square test may be used as a test of independence (comparing frequencies of one nominal variable for different values of a second nominal variable). The chi-square test of independence is used when there are two nominal variables, each of them with two or more possible values. A data set like this is often called an $R \times C$ table, where $R$ is the number of rows and $C$ is the number of columns. In this paper, the two nominal variables are the 6 levels into which the errors are classified (PN, SP, WG, PG, CG and LX), and the specific error description within each level which is compatible across levels (three types of errors: the distinction IT vs. ER, the linguistic function or category affected in the error, and the part of speech in which the error type occurs). The error description of the annotation system allows for finer error types than the above, but these types do not always occur throughout all the error levels (e.g. errors involving the apostrophe occur only in PN, while errors involving phrase heads or dependents occur only in PG), and are therefore not used here. In general, the null hypothesis is that the relative proportions of one variable are independent of the second variable, i.e. that the proportions at one variable are the same for different values of the second variable. In our case, the null hypothesis is that the relative proportions of each of the above types of error (IT vs. ER, linguistic function or category, and part of speech) are the same for each of the 6 error levels.

The Kendall rank correlation coefficient, more commonly referred to as Kendall's tau ($\tau$) coefficient or a tau test, is a non-parametric statistic used to measure the association or statistical dependence between two measured quantities. Specifically, it is a measure of rank correlation, i.e. the similarity of the orderings of the data when ranked by each of the quantities. In our case, this statistic is used for associations between: i) the frequency of errors MS, OM, OR, and OV across the levels PN, SP, WG, PG, CG and LX, and ii) the frequency of errors of different word classes across the levels WG and LX. Kendall's tau ($\tau$) coefficient can be used to explore associations between the variables in a multivariate dataset. We use Kendall's tau ($\tau$) coefficient because it enables to have desirable properties (e.g. robustness) compared to other correlations (e.g. Spearman rank correlation or Pearson correlation). Kendall’s tau represents a probability, i.e. the difference between the probability that two variables are in the same order in the observed data vs. the probability that the two variables are in different orders. In Kendall’s tau ($\tau$) coefficient, if the agreement between the two rankings is perfect (i.e. the two rankings are the same) the coefficient has value 1. If the disagreement between the two rankings is perfect (i.e. one ranking is the reverse of the other) the coefficient has value $-1$. For all other arrangements the value lies between $-1$ and $1$. Increasing values imply increasing agreement between the rankings. If the rankings are completely independent, the coefficient has value 0 on average.

Two tests are used for the same set of data because the data available do not always comply with the assumptions of the chi-square test. A similar test suitable for our data, Kendall’s tau ($\tau$) coefficient, is thus used additionally.
3. Results

The errors found show clear profiles. Thus, PN features 387 errors distributed over 20 types of errors, 6 of which amount to 92.51% of the total. Specifically, three of these (70.54%) involve errors in the use of commas (PN.CM):

<table>
<thead>
<tr>
<th>Error tag</th>
<th>Tag gloss</th>
<th>Absolute value</th>
<th>Percentage</th>
<th>Accumulated percentage</th>
</tr>
</thead>
<tbody>
<tr>
<td>PN.CM.MS</td>
<td>Punctuation, Comma, Misselection</td>
<td>45</td>
<td>11.63%</td>
<td></td>
</tr>
<tr>
<td>PN.CM.OM</td>
<td>Punctuation, Comma, Omission</td>
<td>172</td>
<td>44.44%</td>
<td>70.54%</td>
</tr>
<tr>
<td>PN.CM.OV</td>
<td>Punctuation, Comma, Overinclusion</td>
<td>56</td>
<td>14.47%</td>
<td></td>
</tr>
<tr>
<td>PN.EP.OV</td>
<td>Punctuation, Ellipsis points, Overinclusion</td>
<td>38</td>
<td>9.82%</td>
<td></td>
</tr>
<tr>
<td>PN.FS.OM</td>
<td>Punctuation, Full stop, Omission</td>
<td>21</td>
<td>5.43%</td>
<td></td>
</tr>
<tr>
<td>PN.QR.OV</td>
<td>Punctuation, Quotation marks, Overinclusion</td>
<td>26</td>
<td>6.72%</td>
<td></td>
</tr>
</tbody>
</table>

Examples of the errors in Table 1 follow:

3. I met people and nice teachers &lt;PN.CM.MS&gt;&lt;/PN.CM.MS&gt; we learn grammar, vocabulary and […]

4. We have two legs for walk around the world and see new things and &lt;PN.CM.OM&gt;&lt;/PN.CM.OM&gt; if we can speak different languages &lt;PN.CM.OM&gt;&lt;/PN.CM.OM&gt; we'll learn different cultures.

5. We need to know other languages in order to communicate ourselves with other people, in order to understand each other, in order to negotiate, to establish diplomatic relations with other countries &lt;PN.CM.OV&gt;&lt;/PN.CM.OV&gt; …knowing languages is even important for making war.

6. We need a solution right now &lt;PN.EP.OV&gt;&lt;/PN.EP.OV&gt; it's the same that the solution was given by U.N. or E.U. […]

7. On the other hands, all it's not as easy as we think &lt;PN.FS.OM&gt;&lt;/PN.FS.OM&gt; there're obstacles for example, to have to separate to your family, […]

8. Now you can click with the &lt;PN.QR.OV&gt;"mouse"&lt;/PN.QR.OV&gt; in any chat and you can speak with 25 persons of different countries […]

![Figure 1. Frequency distribution of Punctuation (PN) errors](image-url)
Spelling features 350 errors distributed over 35 types, 6 of which amount to 60.86% of the total. Specifically, 4 of these (37.43%) relate to single graphemes, (SP.GR.SN) and 2 (23.43%) to orthographic case (SP.OS):

Table 2. Main error types in Spelling (SP)

<table>
<thead>
<tr>
<th>Error tag</th>
<th>Tag gloss</th>
<th>Absolute value</th>
<th>Percentage</th>
<th>Accumulated percentage</th>
</tr>
</thead>
<tbody>
<tr>
<td>SP.GR.SN.CN.GL.MS</td>
<td>Spelling, Grapheme, Single, Consonant, General lettering, Misselection</td>
<td>19</td>
<td>5.43%</td>
<td>37.43%</td>
</tr>
<tr>
<td>SP.GR.SN.CN.GL.OM</td>
<td>Spelling, Grapheme, Single, Consonant, General lettering, Omission</td>
<td>66</td>
<td>18.86%</td>
<td></td>
</tr>
<tr>
<td>SP.GR.SN.VW.GL.OM</td>
<td>Spelling, Grapheme, Single, Vowel, General lettering, Omission</td>
<td>20</td>
<td>5.71%</td>
<td></td>
</tr>
<tr>
<td>SP.GR.SN.VW.GL.MS</td>
<td>Spelling, Grapheme, Single, Vowel, General lettering, Misselection</td>
<td>26</td>
<td>7.43%</td>
<td></td>
</tr>
<tr>
<td>SP.OS.LC.PA.MS</td>
<td>Spelling, Orthographic case, Lower case, Provenance Adjective, Misselection</td>
<td>48</td>
<td>13.71%</td>
<td>23.43%</td>
</tr>
<tr>
<td>SP.OS.UC.MS</td>
<td>Spelling, Orthographic case, Upper case, Misselection</td>
<td>34</td>
<td>9.71%</td>
<td></td>
</tr>
</tbody>
</table>

Examples of the errors in Table 2 follow:

(9) […] easy and <SP.GR.SN.CN.GL.MS>comfortable</SP.GR.SN.CN.GL.MS> […]
(10) […] the <SP.GR.SN.CN.GL.OM>tecnology</SP.GR.SN.CN.GL.OM> […]
(11) […] the <SP.GR.SN.VW.GL.OM>turism</SP.GR.SN.VW.GL.OM> […]
(12) […] just like <SP.GR.SN.VW.GL.MS>meny</SP.GR.SN.VW.GL.MS> others […]
(13) […] I had another <SP.OS.LC.PA.MS>english</SP.OS.LC.PA.MS> teacher […]
(14) […] so <SP.OS.UC.MS>In</SP.OS.UC.MS> my opinion […]

Figure 2. Frequency distribution of Spelling (SP) errors
Word grammar features 276 errors distributed over 62 types, 6 of which amount to 44.57% of the total. Specifically, two of these (19.20%) relate to number in nouns (WG.NN.NB) and three (21.74%) to various grammatical categories in verbs (WG.VR):

Table 3. Main error types in Word grammar (WG)

<table>
<thead>
<tr>
<th>Error tag</th>
<th>Tag gloss</th>
<th>Absolute value</th>
<th>Percentage</th>
<th>Accumulated percentage</th>
</tr>
</thead>
<tbody>
<tr>
<td>WG.NN.NB.LU.IT.MS</td>
<td>Word grammar, Noun, Number, Plural, Internal Misselection</td>
<td>15</td>
<td>5.43%</td>
<td>19.20%</td>
</tr>
<tr>
<td>WG.NN.NB.SG.ER.MS</td>
<td>Word grammar, Noun, Number, Plural, External, Misselection</td>
<td>38</td>
<td>13.77%</td>
<td></td>
</tr>
<tr>
<td>WG.PO.DM.NB.SG.ER.MS</td>
<td>Word grammar, Pronoun, Demonstrative, Number, Singular, External, Misselection</td>
<td>10</td>
<td>3.62%</td>
<td></td>
</tr>
<tr>
<td>WG.VR.RS.3P.ER.MS</td>
<td>Word grammar, Verb, Person, Third person singular present indicative, External, Misselection</td>
<td>20</td>
<td>7.25%</td>
<td>21.74%</td>
</tr>
<tr>
<td>WG.VR.RS.BF.ER.MS</td>
<td>Word grammar, Verb, Person, Base form, External, Misselection</td>
<td>24</td>
<td>8.70%</td>
<td></td>
</tr>
<tr>
<td>WG.VR.TN.RT.ER.MS</td>
<td>Word grammar, Verb, Tense, Present, External, Misselection</td>
<td>16</td>
<td>5.80%</td>
<td></td>
</tr>
</tbody>
</table>

Examples of the errors in Table 3 follow:

(15) [...] and <WG.NN.NB.LU.IT.MS>womens</WG.NN.NB.LU.IT.MS> that lived [...]
(16) [...] took my clothes, washed my <WG.NN.NB.SG.ER.MS>tooth</WG.NN.NB.SG.ER.MS>, took my things [...]
(17) [...] thanks to <WG.PO.DM.NB.SG.ER.MS>this</WG.PO.DM.NB.SG.ER.MS> books [...]
(18) [...] there are 3 students, that always <WG.VR.RS.3P.ER.MS>eats</WG.VR.RS.3P.ER.MS> with us [...]
(19) [...] many people say that everybody <WG.VR.RS.BF.ER.MS>are</WG.VR.RS.BF.ER.MS> Spanish [...]
(20) Yesterday I <WG.VR.TN.RT.ER.MS>get</WG.VR.TN.RT.ER.MS> up at 7:40 [...]

Figure 3. Frequency distribution of Word grammar (WG) errors
Phrase grammar features 306 errors distributed over 41 types, three of which amount to 53.92% of the total. Specifically, 2 of these (48.37%) relate to determiner occurrence (PG.CS.DT.ON):

Table 4. Main error types in Phrase grammar (PG)

<table>
<thead>
<tr>
<th>Error tag</th>
<th>Tag gloss</th>
<th>Absolute value</th>
<th>Percentage</th>
<th>Accumulated percentage</th>
</tr>
</thead>
<tbody>
<tr>
<td>PG.CS.CP.VR.RE.PE.OV</td>
<td>Phrase grammar, Constituent, Complement, Verb, Realisation, Preposition, Overinclusion</td>
<td>17</td>
<td>5.56%</td>
<td></td>
</tr>
<tr>
<td>PG.CS.DT.ON.OM</td>
<td>Phrase grammar, Constituent, Determiner, Occurrence, Omission</td>
<td>62</td>
<td>20.26%</td>
<td>48.37%</td>
</tr>
<tr>
<td>PG.CS.DT.ON.OV</td>
<td>Phrase grammar, Constituent, Determiner, Occurrence, Overinclusion</td>
<td>86</td>
<td>28.10%</td>
<td></td>
</tr>
</tbody>
</table>

Examples of the errors in Table 4 follow:

(21) [...] if you regret <PG.CS.CP.VR.RE.PE.OV>of</PG.CS.CP.VR.RE.PE.OV> being there […]
(22) On the other hand, using <PG.CS.DT.ON.OM></PG.CS.DT.ON.OM> English language […]
(23) I went to <PG.CS.DT.ON.OV>the</PG.CS.DT.ON.OV> class 8 […]

Clause grammar features 190 errors distributed over 32 types, three of which amount to 51.58% of the total and relate to the error types described below.
Table 5. Main error types in Clause grammar (CG)

<table>
<thead>
<tr>
<th>Error tag</th>
<th>Tag gloss</th>
<th>Absolute value</th>
<th>Percentage</th>
<th>Accumulated percentage</th>
</tr>
</thead>
<tbody>
<tr>
<td>CG.CS.AB.ON.OR</td>
<td>Clause grammar, Constituent, Adverbial, Occurrence, Ordering</td>
<td>51</td>
<td>26.84%</td>
<td></td>
</tr>
<tr>
<td>CG.CS.DO.RE.NF.MS</td>
<td>Clause grammar, Constituent, Direct object, Realisation, Non-Finite, Misselection</td>
<td>17</td>
<td>8.95%</td>
<td></td>
</tr>
<tr>
<td>CG.CS.SB.ON.OM</td>
<td>Clause grammar, Constituent, Subject, Occurrence, Omission</td>
<td>30</td>
<td>15.79%</td>
<td></td>
</tr>
</tbody>
</table>

Examples of the errors in Table 5 follow:

(24) The problems that affect <CG.CS.AB.ON.OR>now</CG.CS.AB.ON.OR> internet [...]
(25) If the decide <CG.CS.DO.RE.NF.MS>going</CG.CS.DO.RE.NF.MS> out of our country they may enter [...]
(26) [...] because I think <CG.CS.SB.ON.OM></CG.CS.SB.ON.OM> is something very important [...]

Figure 5. Frequency distribution of Clause grammar (CG) errors

Lexis features 303 errors distributed over 52 types, 5 of which amount to 45.54% of the total. Specifically, 2 of these (18.15%) relate to external errors in nouns (LX.NN.ER.MN) and 2 (17.82%) relate to errors in verbs (LX.VR):
Table 6. Main error types in Lexis (LX)

<table>
<thead>
<tr>
<th>Error tag</th>
<th>Tag Gloss</th>
<th>Absolute value</th>
<th>Percentage</th>
<th>Accumulated percentage</th>
</tr>
</thead>
<tbody>
<tr>
<td>LX.NN.ER.MN.CE.MS</td>
<td>Lexis, Noun, External, Meaning, Categorial, Misselection</td>
<td>22</td>
<td>7.26%</td>
<td>18.15%</td>
</tr>
<tr>
<td>LX.NN.ER.MN.LL.US.MS</td>
<td>Lexis, Noun, External, Meaning, Lexical, Usage, Misselection</td>
<td>33</td>
<td>10.89%</td>
<td>29.04%</td>
</tr>
<tr>
<td>LX.PE.ER.MN.LL.US.MS</td>
<td>Lexis, Preposition, External, Meaning, Lexical, Usage, Misselection</td>
<td>29</td>
<td>9.57%</td>
<td>38.61%</td>
</tr>
<tr>
<td>LX.VR.ER.MN.LL.US.MS</td>
<td>Lexis, Verb, External, Meaning, Lexical, Usage, Misselection</td>
<td>27</td>
<td>8.91%</td>
<td>47.52%</td>
</tr>
<tr>
<td>LX.VR.IT.CC.MS</td>
<td>Lexis, Verb, Internal, Collocation, Misselection</td>
<td>27</td>
<td>8.91%</td>
<td>56.43%</td>
</tr>
</tbody>
</table>

Examples of the errors in Table 6 follow:

(27) […] it will be very <LX.NN.ER.MN.CE.MS>importance</LX.NN.ER.MN.CE.MS> because […]
(28) […] people who have studied a specific <LX.NN.ER.MN.LL.US.MS>carrier</LX.NN.ER.MN.LL.US.MS>. for instance people who are dedicated to languages, […]
(29) […] we can read the original books <LX.PE.ER.MN.LL.US.MS>of</LX.PE.ER.MN.LL.US.MS> important foreign writers like shakespeare […]
(30) […] some person <LX.VR.ER.MN.LL.US.MS>have</LX.VR.ER.MN.LL.US.MS> more lucky than other […]
(31) […] you <LX.VR.IT.CC.MS>pass many time</LX.VR.IT.CC.MS> in internet […]

Figure 6. Frequency distribution of Lexis (LX) errors

Figures 1 to 6 show that error distribution varies widely between levels. In PN three error types, which involve use of one and the same item, amount to 70.54% of all errors, and 6 error types amount to 92.51% of all the errors in the level (see Table 1). By contrast, in LX the most frequent error type amounts to 18.15% and the 6 most frequent errors amount to 45.54% of all the errors. Other distribution patterns occur in the rest of levels.

---

6 This category covers errors in categorial meaning, i.e. part of speech. It contrasts with Lexical (LL), which is for errors in lexical meaning.
7 The student may have meant ‘university degree or studies’, but s/he is probably misled by Spanish ‘carrera’ (meaning ‘university degree’).
Across levels, we investigate three parameters: i) executional vs. use errors (IT vs. ER) across 4 levels (WG, PG, CG and LX; SP and PN do not allow the opposition IT vs. ER), ii) surface structure modifications (MS, OM, OR and OV) across the 6 levels of linguistic description, and iii) part of speech across two levels (WG and LX; SP and PN do not allow classification by part of speech, and PG and CG do not for the same categories as WG and LX do). In the first parameter the chi-square analysis shows that the frequencies of IT and ER are significantly different in the levels considered ($\chi^2 = 36.8$, df= 3, $p<0.001$), i.e. the type of error is associated with the type of level. Specifically, the observed frequency of IT errors in WG is lower than the expected frequency, while the observed frequency of ER errors in WG is higher than the expected frequency. Also, the observed frequency of IT errors in CG is remarkably higher than the expected frequency, while the observed frequency of ER errors in CG is remarkably lower than the expected frequency. The observed frequencies and the expected frequencies of IT and ER in LX are similar. The low n of PG may invite questioning these results. Yet, the values without PG, i.e. based on WG, CG and LX, are similar ($\chi^2 = 32.1$, df= 2, $p<0.001$). In the second parameter, the associations among the series MS, OM, OR, and OV across all levels were analyzed with the Kendall rank correlation coefficient:

Table 7. Associations in the series MS, OM, OR and OV according to the Kendall rank correlation coefficient and significance of the statistic

<table>
<thead>
<tr>
<th>Pairs</th>
<th>MS, OR</th>
<th>MS, OM</th>
<th>MS, OV</th>
<th>OR, OM</th>
<th>OR, OV</th>
<th>OM, OV</th>
</tr>
</thead>
<tbody>
<tr>
<td>Tau</td>
<td>-0.60</td>
<td>-0.47</td>
<td>-0.47</td>
<td>0.15</td>
<td>0.15</td>
<td>1.0</td>
</tr>
<tr>
<td>P</td>
<td>0.10</td>
<td>0.27</td>
<td>0.27</td>
<td>0.69</td>
<td>0.69</td>
<td>0.0027</td>
</tr>
</tbody>
</table>

These results show that OM and OV are in perfect agreement (1.0), i.e. they tend to occur in correlation in that their frequencies appear from highest to lowest across the 6 levels in the same order: OM and OV are the highest in PN, then in PG, then in SP, then in CG, then in LX, and the lowest occurs in WG. In contrast, MS and OR are in remarkable disagreement, that is significant only at $p=0.10$, due to the low sample size (low n of levels). An error classification into more levels may result in more significant results. The remaining pairs show tau values close to 0 (i.e. they are independent of each other), even though it would be convenient to increase the sample size.

In the third parameter, part of speech, the chi-square analysis shows significantly different frequencies in the levels WG and LX ($\chi^2 = 126.0$, df= 8, $p<0.001$), i.e. the type of error is associated with the type of level. The Kendall rank correlation coefficient coincides with the chi-square analysis in that there is no correlation between errors according to part of speech with respect to WG and LX (Tau: 0.22, $p=0.39$). The major findings here are higher frequency than expected in WG.AX (Word Grammar, Auxiliary) and WG.PO (Word Grammar, Pronoun), and lower frequency than expected in WG.PE (Word Grammar, Preposition) and, by contrast, higher frequency than expected in LX.PE (Lexis, Preposition) and lower frequency than expected in LX.AX (Lexis, Auxiliary) and LX.PO (Lexis, Pronoun). There is no similarity in the ordering of the frequency of errors of parts of speech in WG and LX.

4. Conclusion

This paper stresses the value of gaining awareness of error distribution and association. The sample shows a clear profile as regards number and type of errors. This profile reveals difficulties, in particular, with the use of commas: the students use them for other punctuation marks, they use them too often and, especially, they leave them out when they should not. Vocabulary shows the opposite picture: the students find difficulties in a larger number of issues and, while some may be dominant (e.g. selection of wrong nouns with regard to lexical and categorial meaning, and of wrong verbs with regard to lexical meaning), there are not as many difficulties as compared to those related to commas in PN. The result is a general difficulty in the use of vocabulary, which is harder to solve than the issue of commas. The rest of levels show a mixed picture between higher dominance in a narrow range of errors in PN and lower dominance in a wider range of errors in LX. SP shows a narrower range of difficulties than PN, but a wider one than the rest. The students find trouble to spell out words in general, even in inflectionally or derivationally-unrelated contexts. They also face problems in selecting low and, especially, upper case. Other difficulties can be found in the rest of levels: in WG, tense and person in verbs and number of nouns; in PG, determiners (left out or included where not appropriate); in CG, the position of adverbials and the lack of a clause.
subject. Finer distinctions than the ones presented here are possible. As to the type of error, the students find more
difficult to use inflections in appropriate contexts (WG.ER) than to select the right allomorphs or relevant
grammatical categories (WG.IT). Conversely, in CG the students find more difficult to construct structures
responding to syntactic processes (CG.IT) than to use them (CG.ER). The students also experience difficulties with
appropriate inclusion or omission of items, and OM and OV are thus correlated throughout all the levels of error
description. In terms of part of speech, auxiliaries and pronouns stand out as containing errors frequently as regards
WG and infrequently as regards LX, i.e. the students know the stock of auxiliary verbs and pronouns, but have
problems inflecting them. This is remarkable in auxiliary verbs, for the low complexity of the inflectional system of
English verbs, and for the high frequencies observed in auxiliaries (AX). Accordingly, work on lexis, which has a
large number of errors, will not have such a widespread effect (in relation to its level of description) as work on the
use of commas. This type of evidence helps decide on the best possible approach and degree of attention to each
issue in ELT.

References

Campoy-Cubillo, M.C., Bellés-Fortuño, B., & Gea Valor, M. L. (Eds.) (2010). *Corpus-based approaches to English
teaching.* London: Continuum.

512-528.


statistical techniques and web search to correct ESL errors. *CALICO, 26*(3), 491-511.

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

8(1), 83-109.


awareness and promote proficiency. Paper presented at the *TALC6 Conference.* University of Granada, Granada,
6-9 July 2004.