The Cambridge Learner Corpus - error coding and analysis for lexicography and ELT

Diane Nicholls
Cambridge University Press (freelance)
diane.nicholls3@btopenworld.com

Abstract

The Cambridge Learner Corpus is a 16 million-word corpus of Learner English collected by Cambridge University Press in collaboration with the University of Cambridge Local Examinations Syndicate (now Cambridge ESOL). It comprises English examination scripts, transcribed retaining all errors, written by learners of English with 86 different mother tongues. The scripts range across 8 EFL examinations and cover both general and business English. A 6 million-word component of the corpus has been error coded to date, using an error-coding system devised at CUP specifically for the Cambridge Learner Corpus. The majority of codes are based on a two-letter system in which the first letter represents the *general type of error* (e.g. wrong form, omission), while the second letter identifies the *word class of the required word.* There are 88 possible codes in all.

This paper will describe the coding system and the corpus tools used for analysis of the coded corpus, and will demonstrate the benefits which this coding and analysis provides for both lexicographers and writers of other ELT books at CUP.

1. Description of the corpus[1]

Since 1993, Cambridge University Press, in collaboration with the University of Cambridge Local Examinations Syndicate (now Cambridge ESOL), has compiled a 16 million-word corpus of Learner English. Students' examination essays are carefully transcribed, reproducing all errors, checked for inputter-generated errors, and stored in the corpus, along with candidate details and examination scores.

The corpus is growing all the time. At present, the complete corpus contains more than 16 million words. 86 mother tongues are represented in the corpus, with more than 350,000 words for more than 15 of the mother tongues. The error-coded component of the corpus currently contains 6 million words.

A profile of each candidate is given for each examination script. This includes information on the first language, age, sex, education history and years of English study of each student. This information can be used to specify the parameters for the creation of subcorpora. For example, it is possible to isolate for analysis the English of very young learners or a particular examination level, mother tongue or language group. A combination of any of these details can be used to create a subcorpus.

2. The coded corpus

6 million words of the Cambridge Learner Corpus have been error coded to date. While error coding is certainly a laborious and time-consuming task, its benefits for our purposes, which I discuss below, have proved to far outweigh the difficulties. The system of error codes and software have been designed in such a way as to overcome, as far as possible, problems with the indeterminacy of some error types. The corpus has also been manually coded by just two coders, with one coder overseeing the work of the second, thus keeping to a minimum any problems with consistency of tagging.

Our aim in coding the corpus is not to create a systematic taxonomy of learner errors but, where possible, to capture under one heading all errors (both of omission and commission) of a particular type so that they can easily be extracted and analyzed and the information gained passed on to examining bodies, teachers, lexicographers, researchers and ELT authors for use in developing tools for learners of English. The codes are not an end in themselves, but rather, act as bookmarks to the contexts in which an error repeatedly occurs. These bookmarks can then be referred to for further analysis.

[1] For further information about the Cambridge Learner Corpus, please refer to links at our website: http://www.cambridge.org/elt/corpus

Also of vital importance is the additional feature of a 'correct' version being added, wherever possible, alongside the error. Care is taken not to 'interpret' or paraphrase errors and only to add a corrected version where there is relative certainty and only one clear replacement possible. This measure vastly increases the search potential of the corpus, allowing access to data which would otherwise be present but only indirectly accessible.

The main advantages of a coded (and 'corrected') corpus over a raw corpus (without error tags) are as follows:

- We do not have to look at everything and scan through both correct and incorrect uses. Correct uses (which are automatically tagged NE (no error)) can be deselected and the remaining cites sorted according to error type for easy look-up and analysis. This also allows the possibility of comparing what learners get right (an often neglected area in ELT) with what they get wrong.

- Using the statistical tools built into our corpus software, we can simply establish the frequency, level of student or mother tongue for a given error (or correct use).

- Perhaps the greatest advantage over an uncoded corpus is that we can search for errors of *omission* as well as *commission*. After searching through a concordanced search on 'at', for example, in an uncoded corpus, it is possible to locate errors such as the unnecessary use of the preposition (e.g. *watching at the television), and the erroneous use of the preposition (e.g. *she invited me at her birthday party). However, it is not possible easily to find:

  i) Instances of failure to use the preposition (coded <#MT>) where it is required (e.g. *we looked each other)

  ii) Instances of where 'at' should have been the chosen preposition, but a wrong preposition was chosen instead (coded <#RT> (e.g. *we arrived to our destination). This would not be possible without the addition of and ability to search on a corrected version.

- An additional feature which is unavailable in a raw corpus, is that we can also search on the error tags themselves, collecting together all errors of a particular type and establishing statistics for their comparative frequencies of occurrence and obtaining information on which students are most likely to make a particular error and at which examination levels. For example, we can look at all noun countability or noun inflection errors in the corpus and establish which are the most common and for which students. This important advantage of the error coded corpus is demonstrated in section 5 below.

- We can also search on clusters of errors. For, example, we can look at all noun-related or adjective-related errors *en masse*.

3. The system of error codes

Learner errors are tagged using the following convention:

<#CODE>*wrong word*|*corrected word*</#CODE>

The majority of the error codes are based on a two-letter coding system in which the first letter represents the *general type of error* (e.g. wrong form, omission), while the second letter identifies the *word class of the required word.*

General types of error (first letter)

F wrong Form used
M something Missing

R	word or phrase needs Replacing
U	word or phrase is Unnecessary (i.e. redundant)
D	word is wrongly Derived

The codes M, R, and U can occur alone where no more specific information can be given.

Word classes (second letter)

A	Pronoun (Anaphoric)
C	Conjunction (linking word)
D	Determiner
J	AdJective
N	Noun
Q	Quantifier
T	Preposition
V	Verb (includes modals)
Y	Adverb (-lY)

Punctuation errors (Error type + P)

Punctuation errors are coded with P as the second letter, and one of the error types M, R, U as the first letter.

MP	punctuation Missing
RP	punctuation needs Replacing
UP	Unnecessary punctuation

When punctuation errors are corrected by the coder, any change in capitalization is also shown within the error coding. For example, the sentence **He died we buried him the next day* omits a full stop after *he died* and therefore requires the <#MP>|.</#MP> coding. However, the new sentence break requires the *w* of *we* to be capitalized. This is shown by the coding *He died* <#MP>*we*|.*We*</#MP> *buried him the next day*. This covers both the punctuation error and the required changes to the capitalization.

Countability errors (C + word class)

CN	countability of Noun error
CQ	wrong Quantifier because of noun countability
CD	wrong Determiner because of noun countability

CN means that the student has used a form which is not available in the intended sense of the noun (e.g. **the country's natural beauties*, **two transports*). Where a noun could be *either* count or uncount but the wrong form has been used, the error is FN, wrong form of noun (e.g. vacation/vacations).

False friend errors (FF + any word class)

All false friend errors are tagged with FF. The word class of the required word is specified by adding a word class code (A, C, D, J, N, Q, T, V, Y) to the FF code. This code is only used when the coder is certain he/she is dealing with a documented False Friend. Otherwise, it is treated as a replace (R) error.

Agreement errors (AG + word class)

AGA	Anaphoric (pronoun) agreement error
AGD	Determiner agreement error
AGN	Noun agreement error
AGV	Verb agreement error

Additional error codes

AS	incorrect Argument Structure
CE	Compound Error
CL	CoLlocation error
ID	IDiom error
IN	Incorrect formation of Noun plural
IV	Incorrect Verb inflection
L	inappropriate register (Label)
S	Spelling error
SA	American Spelling
SX	Spelling confusion error
TV	wrong Tense of Verb
W	incorrect Word order
X	incorrect formation of negative

AS (argument structure error) covers errors in argument structures which cannot be coded as MT (missing preposition, e.g. **he explained me*) or UT (unnecessary preposition, e.g. **he told to me*). AS is particularly used for double object verbs, e.g. **it caused trouble to me* is coded <#AS>*it caused trouble to me*|*it caused me trouble*</#AS> to circumvent the need for multiple codes to correct what is, in fact, a single error.

CE (complex error) is a catch-all code to cover multiple errors and groups of words the intended sense of which cannot be established. By using this code, we factor out of the equation strings which can yield little useful information on learner errors.

SA (American Spelling) is not a true *error* code since it is not always possible to know whether a learner has made a mistake or is deliberately using US spelling, but was introduced in view of the fact that learners are usually very clear about which variety of English they wish to learn and ELT and dictionary publishers still need to highlight the differences between the varieties of English in their products. The SX code (spelling confusion error) covers spelling confusables such as to/too, their/there and weather/whether. It is possible to override the distinctions made by these codes by searching on the general-spelling group code <#SPELL>.

4. Some practical issues

4.1. Avoiding over-coding and 'creating' errors

We are not attempting to *rewrite* the scripts into perfect English or to *interpret* the text. Often, things could be expressed better by paraphrase - this is not our task. We are only correcting and documenting errors. Equally, the coder must resist the temptation to make moral judgements about a student's intended meaning. If the language used is 'correct', the idea behind it is not brought into question.

4.2. Embedding errors

In cases where a word is both wrongly spelled and the wrong word, for example, one code can be embedded within another.
e.g. I like to <#S>hospitelize|hospitalise</#S> my relatives - the student means ‘put them up’ or ‘have them to stay’ (show them hospitality). This should be an <#RV> error, not a spelling error and the spelling correction should be embedded inside the more significant RV code. Thus:
<#RV><#S>hospitelize|hospitalise</#S>|put up</#RV>.

4.3. Choosing error codes in ambiguous cases

Coders are occasionally faced with a decision between two different codes. This is the case in, for example: *He said me that ... The coder must decide whether to correct this by replacing the verb (RV) with a synonym (told) which works in the given argument structure (He *told* me that), or to add the missing preposition (MT) to make the argument structure of the student’s chosen verb correct (He said *to* me that).

We want to keep as close as possible to the student’s original text, so it is helpful to think of the error codes as having a loose hierarchy. In this case, changing a verb’s argument structure is a less severe change than changing the verb itself and puts right what the student said wrongly, rather than starting again

from scratch. So, 'He said <#MT>|to</#MT> me that' is more helpful, than, 'He <#RV>said|told</#RV> me that'. It is more helpful to teach students to use a verb correctly than to teach them to avoid using a verb which they use incorrectly. Similarly, where it is impossible to know what the student intended in a Replace error, no correct version is inserted, as a coder's 'wild guess' can be of little help and will distort the data.

5. Using the data

At CUP, the data gathered from the coded corpus is used by lexicographers and researchers working on dictionary projects to identify those words and constructions which are particularly problematic for learners. Based on the corpus information gathered, decisions are made about how best to direct the learner to the correct use of such words or constructions. Information about the language groups affected is also relevant. Attention is also paid to the examination levels at which particular errors occur. This search option is particularly useful as it enables us to establish which errors are typically elementary-level errors, for example, as well as to identify those more 'intransigent' errors which still persist, even at Proficiency level.

Corpus data is also used by course and reference book authors to inform their work and by Cambridge ESOL to inform course designing, analyze marking schemes and to support their on-going examination research.

5.1 Searching on individual words

Figure 1 shows a screen from a search on the word 'Hardly':

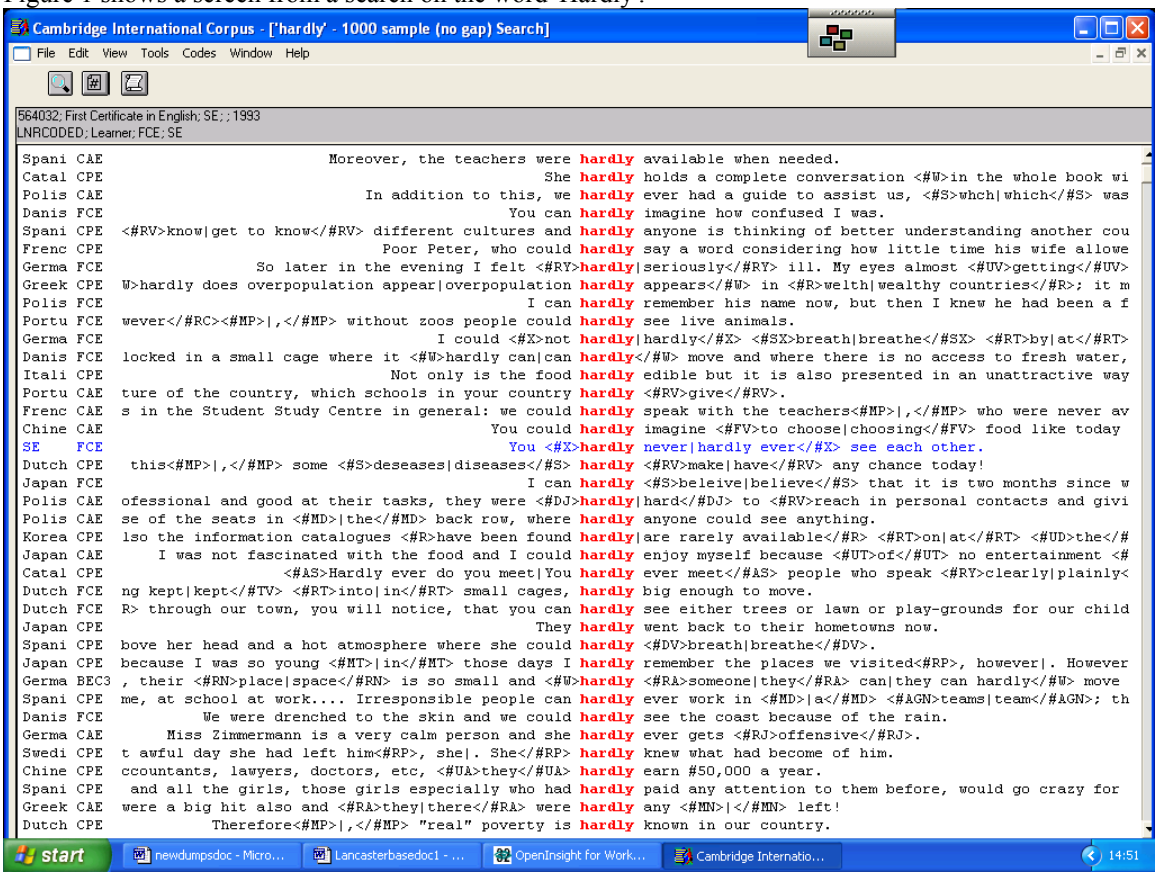

Figure 1

The search is randomly sorted at this stage. The user can scroll up and down the search. As the user scrolls through the search, the blue line (here, in the centre of the screen) indicates the script which the user is currently looking at. The information to the left shows the mother tongue and examination level of the candidate. More candidate information for the relevant script appears in the grey field at the top left hand

corner of the screen and is important for verifying whether an error is idiosyncratic to one individual student or occurs with a number of different students. There is the option to select an entire script for further analysis by moving the cursor to the required script and simply pressing Return.

We can then get the statistics for the error. The statistics for error tags for the search word 'Hardly' are shown in figure 2 below:

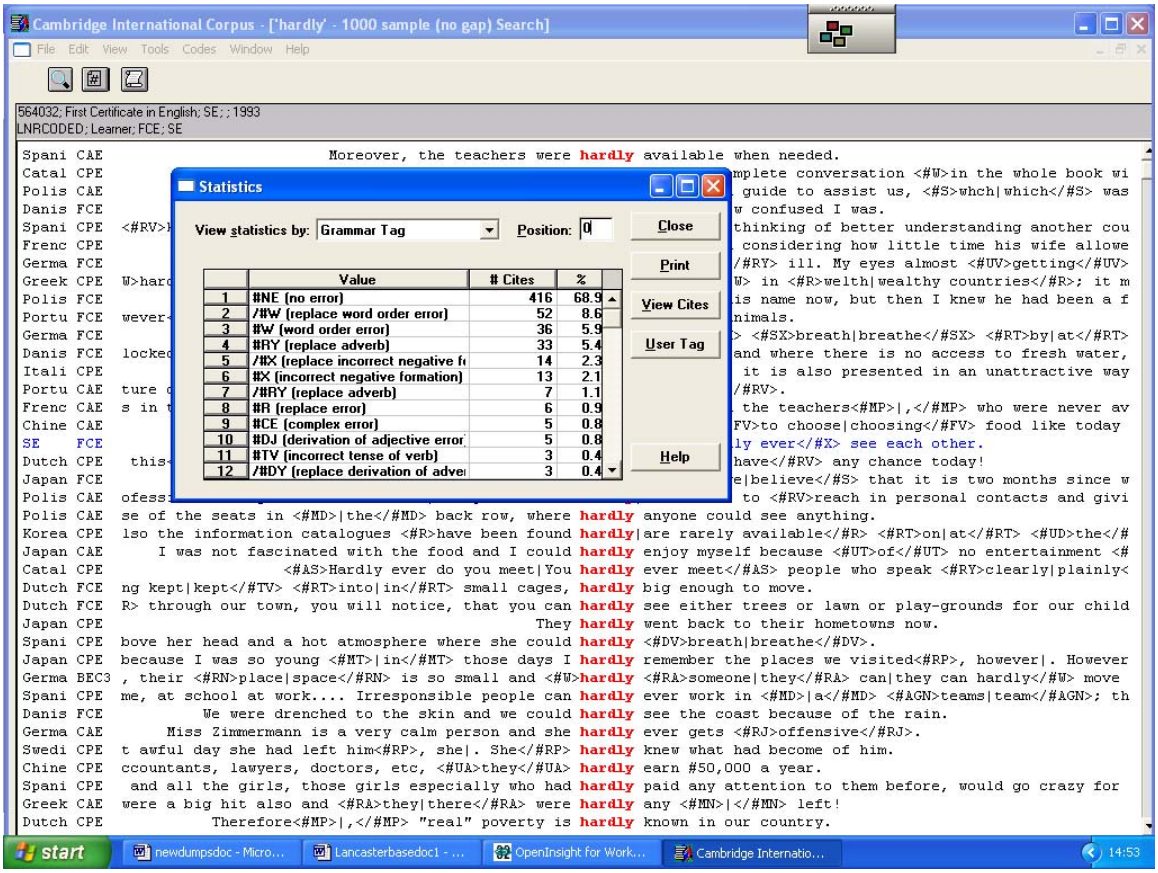

Figure 2

This tells us that the frequency of correct use (NE) is 68.9% (416 citations); that the most frequent error is one of word order (W) (8.6%); the next most frequent error is one of wrong adverb use (RY) (5.4%) and the next, bad negative formation (X) (2.3%). From the statistics box, we can select the error that most interests us and display the citations containing that error. The statistics are not viewed as concrete evidence in themselves and can, admittedly, be distorted by many factors. Rather, they act as a clear signpost to a significantly recurring error and can be used as a path to viewing the errors themselves in context. The final judgment on the importance of an error and how to treat it lies with the analyst/lexicographer.

Figure 3 shows results obtained when selecting the word order error (W):

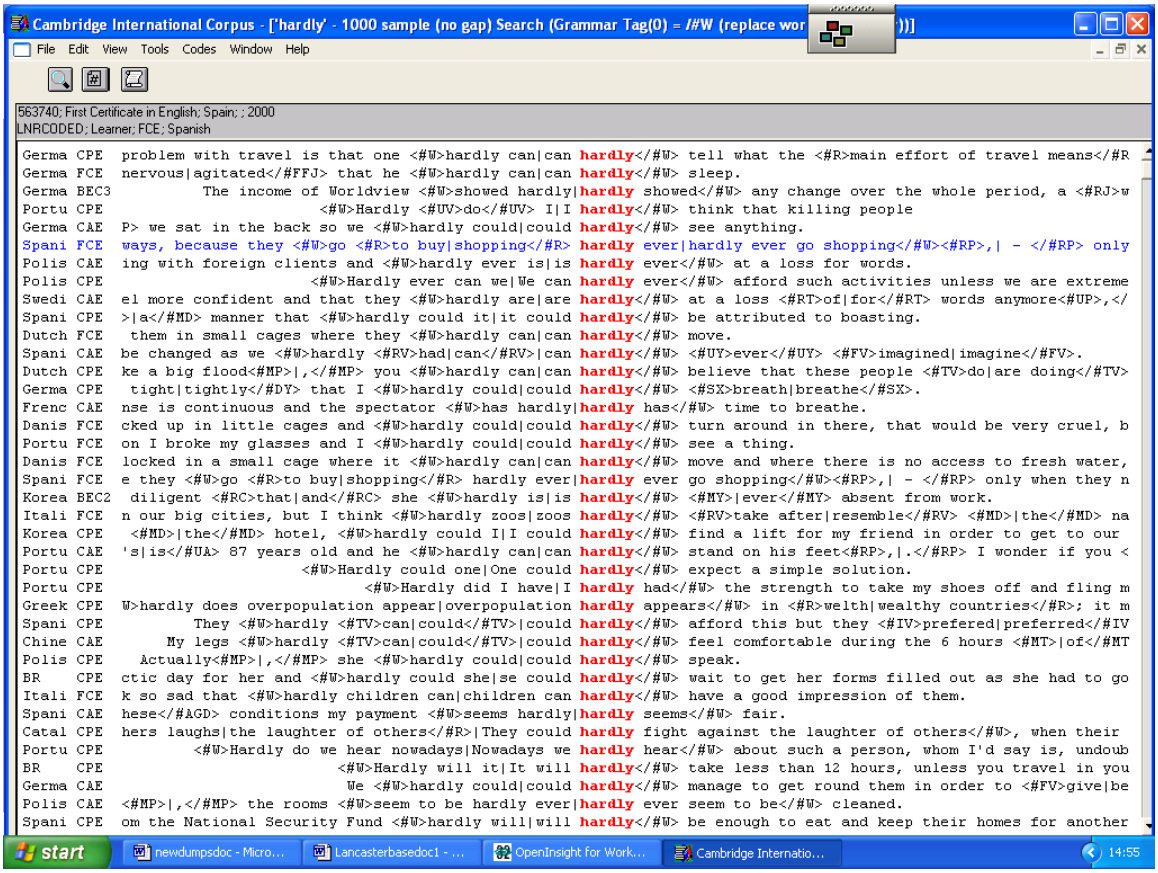

Figure 3

Our search is also not restricted to the search word itself. Using the Position field in the Statistics box, we can select 0 to get statistics for the error tags of the search word. By putting '1' in the Position field, we get the error tags associated with words following the search word. This is very useful for establishing the following environments of a verb, for example. If we search on a verb (e.g. enjoy) and insert '1' in the position field, we find a high percentage of verb form errors (FV). If we select this statistic and View Cites, we find that learners frequently use the infinitive form of a following verb (**I enjoy to swim*) rather than the -ing form (*I enjoy swimming*) with enjoy. It is then a short step to finding which mother tongues make this error most and at which examination levels. To examine preceding environment errors, a '-1' is inserted in the position field.

5.2 Searching on error tags

As discussed in section 3 above, a coded corpus gives the important advantage of being able to search on particular error types or groups of errors in exactly the same way as we search on individual words. For example, we can search for all noun countability errors (CN) in the corpus. Countable nouns are nouns which can take a plural or an indefinite article (e.g. tree), while uncountable nouns cannot take an indefinite article and are never used in the plural (e.g. information). Noun countability is an important area of confusion for learners of English, not least because a noun which is countable in their mother tongue is often inexplicably uncountable in English.

Figure 4 shows a screen from a randomly sorted search on noun countability errors (CN):

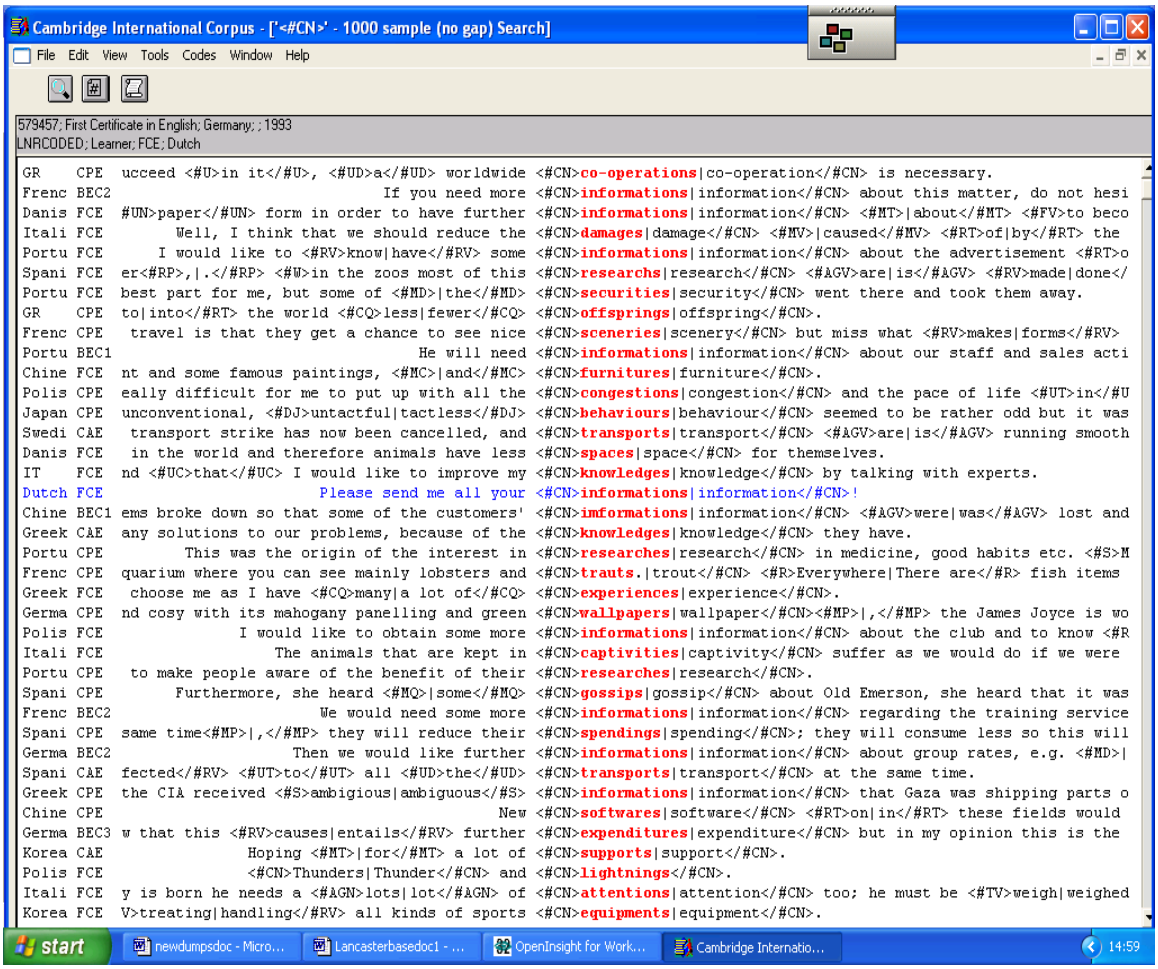

Figure 4

This search can then be sorted for ease of reference and the statistics box used to obtain a list of the words most frequently involved in countability errors:

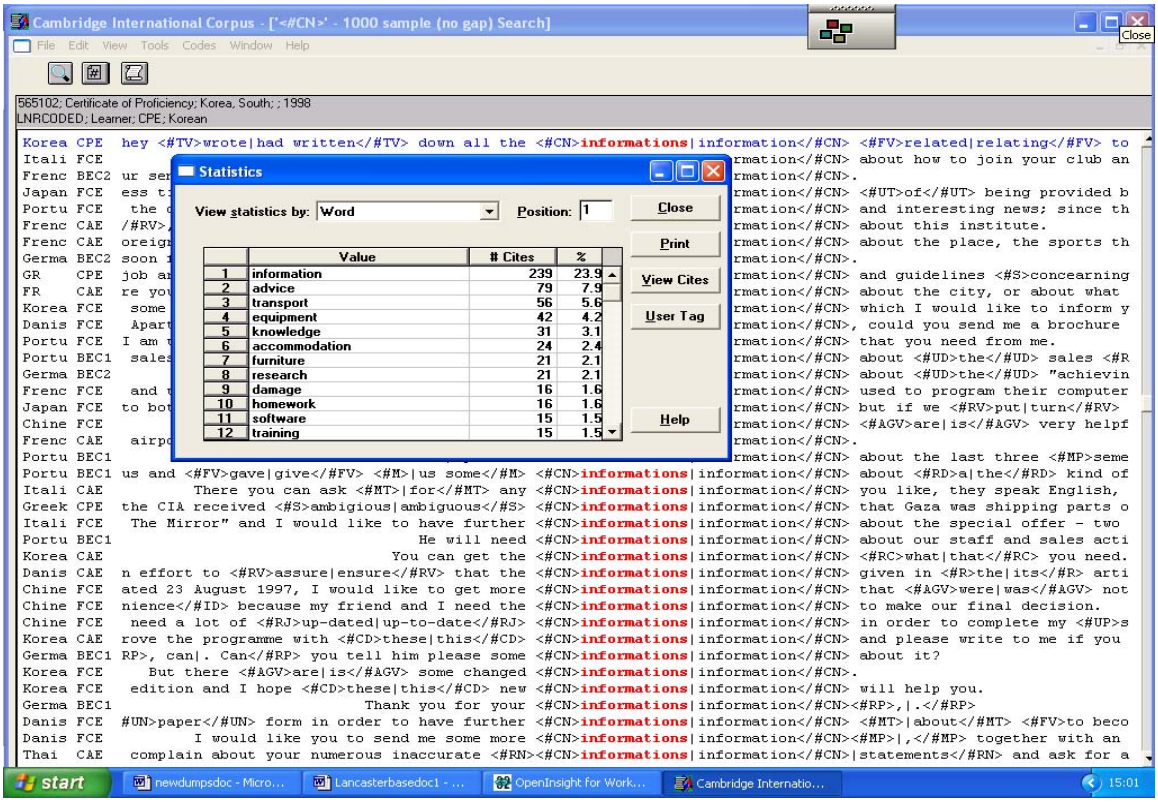

Figure 5

We can also look at the countability errors made by individual mother tongues[2]:

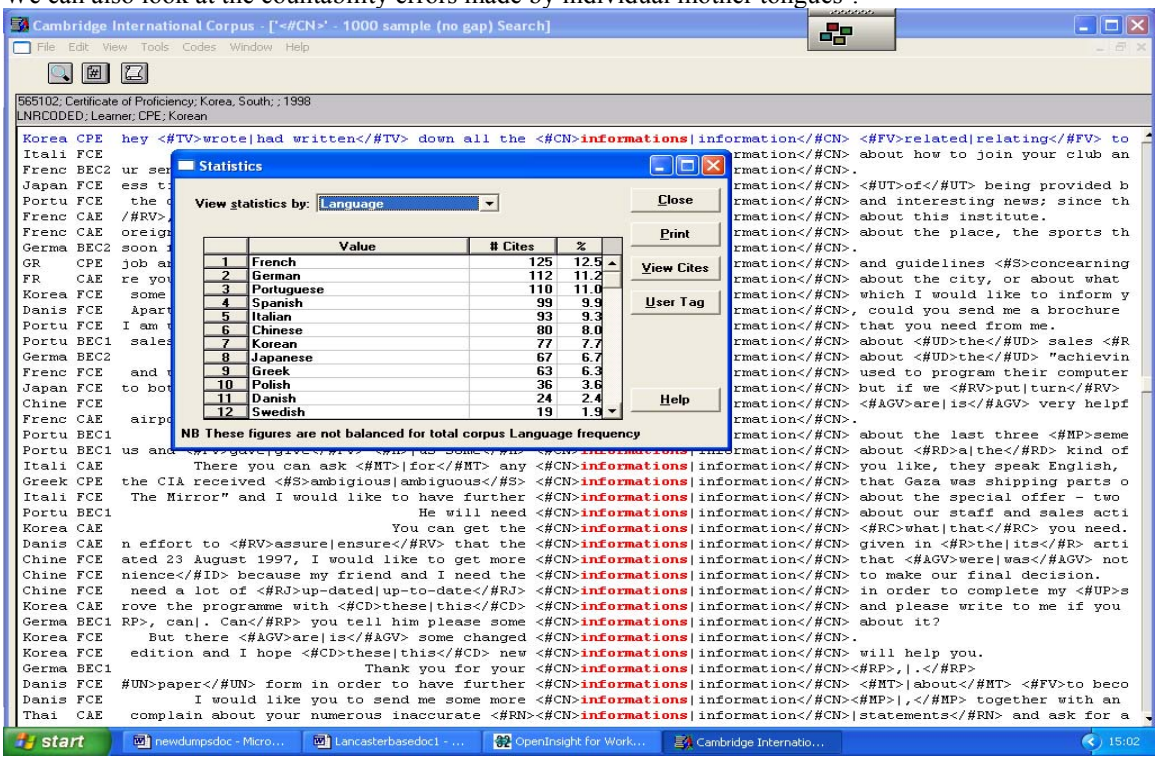

Figure 6

[2] A forthcoming version of the software will be balanced for corpus frequency, enabling us more easily to establish which mother tongues most frequently encounter problems with noun countability or other errors.

We can also obtain information on the examination levels at which these errors occur. These are shown in figure 7 below. It is encouraging to note that the incidence of countability errors gradually decreases as the students progress through examination levels to Proficiency level. However, for teachers and ELT authors alike, the fact that the incidence of error is still quite high (28.2% of these errors are made at Proficiency level) is a clear indication that more needs to be done to alert learners to this major pitfall in English language learning.

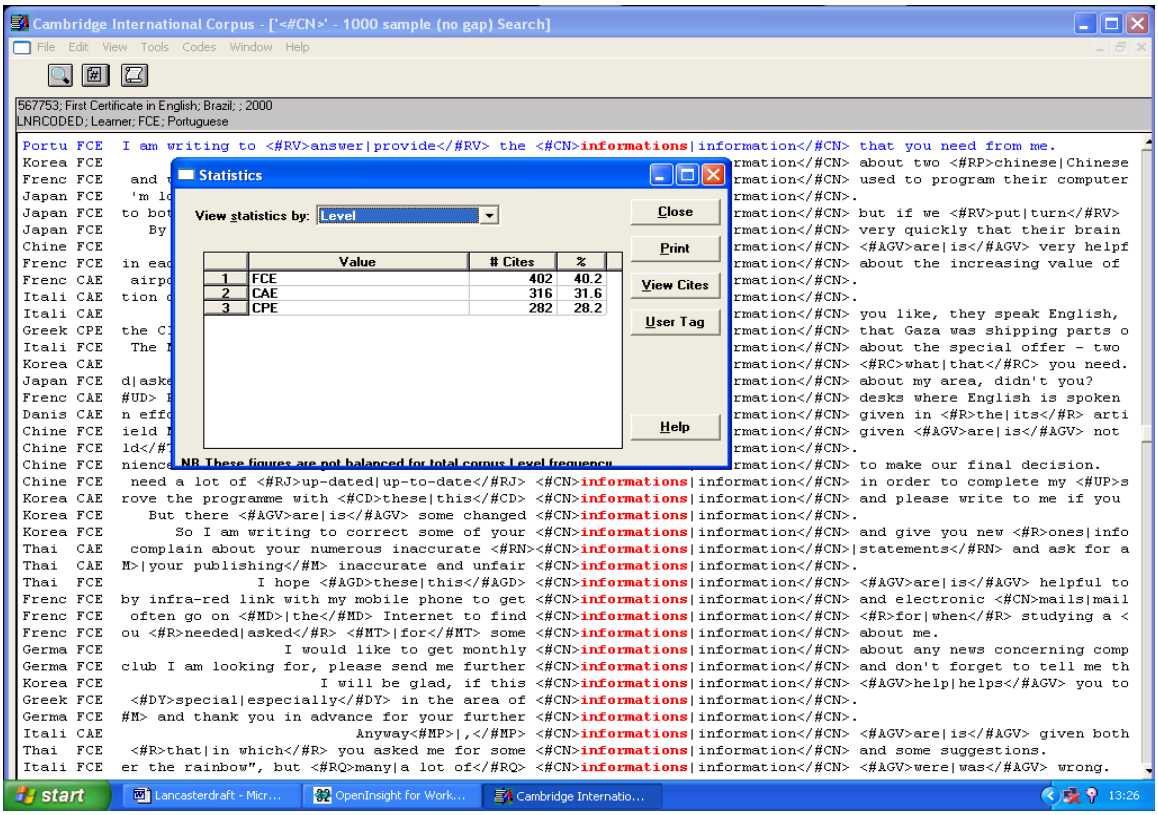

Figure 7

6. Conclusion

Formerly, publishing houses engaged in producing ELT reference and course books were dependent primarily on the intuitions of highly skilled and experienced lexicographers to anticipate learners' difficulties with English. With a learner corpus up and running, these vital human skills are powerfully reinforced and supplemented by a database of 'living' errors and, most importantly, the contexts in which they commonly occur.

The addition of error coding makes it possible to follow a structured, step-by-step procedure to arrive at the required data quickly, efficiently and informatively.

A search on an uncoded corpus must, necessarily, depend largely on the searcher's expectations of what he/she is likely to find, and can only be as fruitful as the searcher is inventive in his/her searching strategies. In addition, with an uncoded corpus, a searcher can only search for what is there and can have only indirect and, to a large extent, 'accidental', access to what is not there - to errors of omission and corrected versions.

The Cambridge Learner Corpus, with the error coding and corpus tools developed for its exploitation, is providing lexicographers, researchers, ELT authors and examiners with easy, direct access to a fund of information which they can interpret and use for widely varying applications.