The Cambridge Learner Corpus - error coding and analysis for lexicography and ELT Diane Nicholls Cambridge University Press (freelance) <u>diane.nicholls3@btopenworld.com</u>

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 corpus¹

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 <u>M</u>issing

- R word or phrase needs <u>R</u>eplacing
- U word or phrase is <u>Unnecessary</u> (i.e. redundant)
- D word is wrongly <u>D</u>erived

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 <u>M</u>issing
- RP punctuation needs <u>R</u>eplacing
- UP <u>Unnecessary punctuation</u>

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 <u>Anaphoric</u> (pronoun) agreement error
- AGD <u>D</u>eterminer agreement error
- AGN <u>N</u>oun agreement error
- AGV <u>V</u>erb agreement error

Additional error codes

- AS incorrect Argument Structure
- CE Compound Error
- CL CoLlocation error
- ID IDiom error
- Incorrect formation of Noun plural IN
- IV Incorrect Verb inflection
- L inappropriate register (Label)
- S Spelling error
- SA American Spelling
- SX Spelling confusion error
- ΤV wrong Tense of Verb
- W incorrect Word order
- Х 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 $\langle \#RV \rangle$ 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':

🛃 Cambridge	International Corpus - ['hardly' - 1000 sample (no gap) Search]
🦳 File Edit Vi	/ew Tools Codes Window Help
564032; First Cert LNRCODED; Lea	tificate in English; SE; ; 1993 amer; FCE; SE
Spani CAE Catal CPE Polis CAE Danis FCE Spani CPE Frenc CPE Germa FCE Germa FCE Dortu FCE Oanis FCE Danis FCE Danis FCE Danis CAE Frenc CAE SE FCE Ducch CPE Ducch CPE Ducch CPE Polis CAE Polis CAE	
Danis FCE Germa CAE Swedi CPE Chine CPE Spani CPE Greek CAE Dutch CPE	<pre>I was not fascinated with the food and I could hardly enjoy myself because <#UT>of<!--#UT--> no entertainment <#</pre>
🦺 start	🛛 🖻 newdumpsdoc - Micro 🕲 Lancasterbasedoc 1 👯 OpenInsight for Work 🚮 Cambridge Internatio



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:

	2 le in English; SE; ; 1993 FCE; SE	
Spani CAE	Moreover, the teachers were hardly available when needed.	
Catal CPE		<#W>in the whole book w
Polis CAE		<#S>whch which #S wa
Danis FCE	w confused I was.	
	The state by araminar rag	nderstanding another co
Frenc CPE		tle time his wife allow
Germa FCE		lmost <#UV>getting #UV</td
		thy countries #R ; it
Polis FCE		n I knew he had been a
	2 that (made and a second 20 E 0	
Germa FCE	4 #PV (replace adverb) 22 54 User Tag	e #SX <#RT>by at #RT</td
	ocked 5 /#X (replace incorrect penative fr 14 2.3 and where there is n	o access to fresh water
tali CPE		d in an unattractive wa
	are c 7 /#RY (replace adverb) 7 1.1 /#RV>.	
		#MP who were never a
Chine CAE		.g #FV food like today
E FCE	11 HTVC SEE CAC	
utch CPE	12 (#DX (replace derivation of advect 2 0.4 - nave #RV any chance	
Japan FCE	e believe #S that	it is two months since
		rsonal contacts and giv
	e of the seats in <#MD> the #MD back row, where hardly anyone could see anything.	
	so the information catalogues <#R>have been found <mark>hardly</mark> are rarely available #RT</td <td></td>	
apan CAE	I was not fascinated with the food and I could hardly enjoy myself because <#UT>of </td <td></td>	
atal CPE	<#AS>Hardly ever do you meet You hardly ever meet #AS people who spe	ak <#RY>clearly plainly
	g kept kept #TV <#RT>into in #RT small cages, hardly big enough to move.	
	> through our town, you will notice, that you can hardly see either trees or lawn or pl	. 2017년 - 17월, GAN (1996) 11일 - 11일 (1997) 11
apan CPE	They hardly went back to their hometowns n	ow.
	ove her head and a hot atmosphere where she could hardly <#DV>breath breathe #DV .	
	ecause I was so young <#MT> in #MT those days I hardly remember the places we visited	
	their <#RN>place space #RN is so small and <#W>hardly <#RA>someone they #RA can th	
	e, at school at work Irresponsible people can hardly ever work in <#MD> a #MD <#A	
anis FCE	We were drenched to the skin and we could hardly see the coast because of the r	
erma CAE	Miss Zimmermann is a very calm person and she hardly ever gets <#RJ>offensive #RJ	
	awful day she had left him<#RP>, she . She #RP hardly knew what had become of him.	
	countants, lawyers, doctors, etc, <#UA>they #UA hardly earn #50,000 a year.	
Spani CPE	and all the girls, those girls especially who had hardly paid any attention to them bef	ore, would go crazy for
	ere a big hit also and <#R&>they there #R& were hardly any <#MN> #MN left!	
utch CPE	Therefore<#MP> , #MP "real" poverty is hardly known in our country.	

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):

🕼 Cambridge International Corpus - ['hardly' - 1000 sample (no gap) Search (Grammar Tag(0) = /#W (replace wor 🌓 1)]
File Edit View Tools Codes Window Help
(563740): First Certificate in English; Spain; ; 2000 LNRCODED ; Learner; FCE ; Spanish
Germa CPE problem with travel is that one <#W>hardly can can hardly #W tell what the <#R>main effort of travel means #R</th
🛃 Starit 📓 newdumpsdoc - Micro 📓 Lancasterbasedoc 1 🙀 OpenInsight for Work 👔 Cambridge Internatio 🔇 14:55

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):

🗿 Cambridge International Corpus - ['<#CN>' - 1000 sample (no gap) Sea	
Tile Edit View Tools Codes Window Help	
579457; First Certificate in English; Germany; ; 1993 LNRCODED; Learner; FCE; Dutch	
GR CPE ucceed <#U>in it #U , <#UD>ac/#UD> worldwide Frenc BEC2 If you need more Danis FCE #UN>paper #UN form in order to have further Itali FCE Well, I think that we should reduce the Portu FCE I would like to <#RV>know have #RV some Spani FCE er<#RP>, . #RP <#W>in the zoos most of this Portu FCE best part for me, but some of <#MD> the #MD GR CPE to into #RT the world <#CQ>less fewer #CV</td Frenc CPE travel is that they get a chance to see nice Portu BEC1 He will need Chine FCE nt and some famous paintings, <#MC> and #MC Polis CPE eally difficult for me to put up with all the Japan CPE unconventional, <#DJ>untactful!tactless #DJ</td Swedi CAE transport strike has now been cancelled, and Datais FCE in the world and therefore animals have less IT <fce< td=""> no <#DVC>that move mail your Chine BEC1 ems broke down so that some of the customers' Greek CAE any solutions to our problems, because of the Portu CPE This was the origin of the interest in Frenc CPE quarium where you can see mainly lobsters and</fce<>	<pre><#CN>sceneries scenery<!--#CN--> but miss what <#RV>makes forms<!--#RV--> <#CN>informations information<!--#CN--> about our staff and sales acti <#CN>furnitures furniture<!--#CN-->. <#CN>furnitures furniture<!--#CN--> and the pace of life <#UT>in<!--#U<br--><#CN>behaviours behaviour<!--#CN--> and the pace of life <#UT>in<!--#U<br--><#CN>behaviours behaviour<!--#CN--> asemed to be rather odd but it was <#CN>transports transport<!--#CN--> <#AGU>are is<!--#AGV--> running smooth <#CN>informations information<!--#CN--> <#AGU>were is<!--#AGV--> lost and <#CN>knowledges knowledge<!--#CN--> by talking with experts. <#CN>informations information<!--#CN--> <#AGV>were was<!--#AGV--> lost and <#CN>knowledges knowledge<!--#CN--> they have. <#CN>researches research<!--#CN--> in medicine, good habits etc. <#S>M <#CN>trants. trout<!--#CN--> <#R>Everywhere There are<!--#R--> fish items <#CN>captivities captivity<!--#CN--> suffer as we would do if we were <#CN>informations information<!--#CN--> captivities captivity<!--#CN--> would be and to know <#R <#CN>captivities captivity<!--#CN--> suffer as we would do if we were <#CN>informations information<!--#CN--> regarding the training service <#CN>esperinges research<!--#CN-->. <#CN>informations information<!--#CN--> regarding the training service <#CN>informations information<!--#CN--> regarding the training service <#CN>informations information<!--#CN--> about group rates, e.g. <#MD> <#CN>informations information<!--#CN--> that Gaza was shipping parts o <#CN>informations information<!--#CN--> the in my opinion this is the <#CN>expenditures spendire<!--#CN-->.</pre>
Korea FCE V>treating handling #RV all kinds of sports # Start Image: start	<#CN>equipments equipment #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:

Cambridge International Corpus - ['<#CN>' - 1000 sample (no	gap) Search]			
File Edit View Tools Codes Window Help				Close
565102; Certificate of Proficiency; Korea, South; ; 1998 LNRCODED : Learner: CPE : Korean				
Itali FCE	all the <#CN>infor			t how to join your club an
Frenc BEC2 ur ser Statistics			rmation #CN .	a tea an Anne Anne Anne ann
Japan FCE ess t				>of #UT being provided b
Portu FCE the View statistics by: Word	▼ Position: 1	Close		interesting news; since th
Frenc CAE /#RV>, Frenc CAE oreign			rmation #CN abou	t this institute. t the place, the sports th
Germa BEC2 soon 1 Value	# Cites %	Print	rmation #CN abou	t the place, the sports th
GR CPE job at 1 information	239 23.9	View Cites		quidelines <#S>concearning
FR CAE re you 2 advice	79 7.9	View Cites		t the city, or about what
Korea FCE some 3 transport 4 equipment	56 5.6 42 4.2	User Tag	rmation #CN whic	h I would like to inform y
Danis FCE Apart 5 knowledge	42 4.2	Userrag	rmation #CN , cou	ld you send me a brochure
Portu FCE I am t 6 accommodation	24 2.4		rmation #CN that	
Portu BEC1 sales 7 furniture	21 2.1			t <#UD>the #UD sales <#R
Germa BEC2 8 research	21 2.1			t <#UD>the #UD "achievin
Frenc FCE and t 9 damage	16 1.6 16 1.6			to program their computer
Japan rez co sou	15 1.5	Help		if we <#RV>put turn #RV
thine FCE 12 training	15 1.5 -	Terb		V>are is #AGV very helpf
Frenc CAE airpo			rmation #CN .	
Portu BEC1				t the last three <#MP>seme
Portu BEC1 us and <#FV>gave give #FV <#M> us so Itali CAE There you can ask <#MT> for #</td <td></td> <td></td> <td></td> <td></td>				
Greek CPE the CIA received <#S>ambigious ambiguo				
Itali FCE The Mirror" and I would like to have				
				t our staff and sales acti
				>what that #RC you need.
Danis CAE n effort to <#RV>assure ensure #RV t				
Chine FCE ated 23 August 1997, I would like to g				
Chine FCE nience #ID because my friend and I n				
Chine FCE need a lot of <#RJ>up-dated up-to-dat	e #RJ <#CN>infor	mations inf	formation #CN in o	rder to complete my <#UP>s
Korea CAE rove the programme with <#CD>these thi	.s #CD <#CN>infor	mations inf	formation #CN and	please write to me if you
Germa BEC1 RP>, can . Can #RP you tell him plea	ise some <#CN> infor	mations inf	formation #CN abou	t it?
Korea FCE But there <#AGV>are is #AGV some				
Korea FCE edition and I hope <#CD>these this #</td <td></td> <td></td> <td></td> <td></td>				
			formation #CN <#RP>	
Danis FCE #UN>paper #UN form in order to have				
Danis FCE I would like you to send me so				
Thai CAE complain about your numerous inaccura	te <#RN><#CN>INFOR	nations 1hi	cormation #CN stat	ements #RN and ask for a
start 💀 newdumpsdoc - Micro 🕲 Lancasterbasedoc1	. 🛛 🙀 OpenInsight for Wor	k 👔 Cam	bridge Internatio	 15:01
Figure 5				
We can also look at the countability errors	made by indivi	dual mot	her tongues ² .	
we can also look at the countaolity chois	made by marvi	uuai mot	ner iongues .	

🚮 Cambridge In	ternational Corp	pus - ['<#CN>' - 1000 sample (no g	ap) Search]			
File Edit View	/ Tools Codes \	Window Help				_ # ×
	2					
565102; Certificate o	of Proficiency; Korea,	, South; ; 1998				
LNRCODED; Learne	er; CPE; Korean					
Korea CPE h	hey <#TV>wrot	te had written #TV down a	all the <#CN>inform	nations inf	ormation #CN <#FV>re	elated relating #FV to .
Itali FCE	- Income					how to join your club an
Frenc BEC2 u		stics			rmation #CN .	
	ess t:			Close		f #UT being provided b
Portu FCE Frenc CAE /	the d View g	statistics by: Language	-			teresting news; since th
	1				rmation #CN about t	the place, the sports th
Germa BEC2 a	oreign	Value	# Cites %	Print	rmation #CN about (the place, the sports th
	job an 1	French	125 12.5			idelines <#S>concearning
	re you 2		112 11.2	View Cites		the city, or about what
Korea FCE	some 3		110 11.0			I would like to inform y
Danis FCE	Amore 4		99 9.9	User Tag		you send me a brochure
Portu FCE 1	I am t 6		93 9.3		rmation #CN that vo	
Portu BEC1	sales 7		77 7.7		rmation #CN about -	<#UD>the #UD sales <#R
Germa BEC2	8		67 6.7		rmation #CN about -	<#UD>the #UD "achievin
Frenc FCE	and t 9		63 6.3		rmation #CN used to	o program their computer
Japan FCE t		Polish	36 3.6		rmation #CN but if	we <#RV>put turn #RV
Chine FCE	11		24 2.4	Help	rmation #CN <#AGV>a	are is #AGV very helpf
Frenc CAE	airpo				rmation #CN .	
Portu BEC1		se figures are not balanced for total c				the last three <#MP>seme
		gavelgives, wive som				
Itali CAE		≥ you can ask <#MT> for #M</td <td></td> <td></td> <td></td> <td></td>				
Greek CPE t Itali FCE		ived <#S>ambigious ambiguou				
Portu BEC1	The HILLOL"	and I would like to have f				our staff and sales acti
Korea CAE						hat that #RC you need.
	n effort to <	<pre> /#RV>assure ensure<!--#RV--> th</pre>				
		st 1997, I would like to ge				
		because my friend and I ne				
Chine FCE	need a lot o	of <#RJ>up-dated up-to-date	#RJ <#CN>infor	nations inf	ormation #CN in orde	er to complete my <#UP>s
Korea CAE r	rove the prog	gramme with <#CD>these this	s #CD <#CN>inform	mations inf	ormation #CN and ple	ease write to me if you
	Germa BEC1 RP>, can . Can #RP you tell him please some <#CN>informations information #CN about it?					
Korea FCE		<#AGV>are is #AGV some o				
Korea FCE	edition and	I hope <#CD>these this #C</td <td></td> <td></td> <td></td> <td></td>				
Germa BEC1					ormation #CN <#RP>, .	
		UN> form in order to have f				
Danis FCE		uld like you to send me son				
Thai CAE	complain abo	out your numerous inaccurat	C ARRIVE ACINE INTON	acions 1ni	ormacion #CN stateme	encovy#KN> and ask IOr a
	newdumpsdoc	- Micro 🛛 💌 Lancasterbasedoc1	OpenInsight for Worl	🤬 📓 Camb	oridge Internatio	15:02
Figure 6						
-8						

 $^{^{2}}$ 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.

🚮 Cambridge	International Corpus - ['<#CN>' - 1000 sample (no gap) Search]
File Edit V	riew Tools Codes Window Help
	tificate in English; Brazil: ; 2000 amer; FCE; Portuguese
Portu FCE	I am writing to <#RV>answer provide #RV the <#CN>informations information #CN that you need from me.
Korea FCE	mation #CN about two <#RP>chinese Chinese
Frenc FCE	and v = Statistics
Japan FCE	'm lo rmation #CN .
Japan FCE	to bot View statistics by: Level
Japan FCE	By rmation #CN very quickly that their brain
Chine FCE	Print rmation #CN <#AGV>are is #AGV very helpf
Frenc FCE	in eac Value # Cites % rmation #CN about the increasing value of
Frenc CAE	airpo FCE 402 40.2 View Cites rmation #CN .
Itali CAE	tion d 2 CAE 316 31.6 rmation #CN .
Itali CAE	3 CPE 282 28.2 User Tag rmation #CN you like, they speak English,
Greek CPE	the C rmation #CN that Gaza was shipping parts o
Itali FCE	The I rmation #CN about the special offer - two
Korea CAE	rmation #CN <#RC>what that #RC you need.
Japan FCE	d aska rmation #CN about my area, didn't you?
Frenc CAE	#UD> 1 rmation #CN desks where English is spoken
Danis CAE	n effo rmation #CN given in <#R>the its #R arti
Chine FCE	ield I Help rmation #CN given <#AGV>are is #AGV not
Chine FCE	ld #</th
Chine FCE	nience NR These figures are not balanced for total corpus Level frequency rmation #CN to make our final decision.
Chine FCE	need a lot of <#RJ>up-dated up-to-date #RJ <#CN> informations information #CN in order to complete my <#UP>s
Korea CAE	rove the programme with <#CD>these this #CD <#CN> informations information #CN and please write to me if you
Korea FCE	But there <#AGV>are is #AGV some changed <#CN>informations information #CN .
Korea FCE	So I am writing to correct some of your <#CN> informations information #CN and give you new <#R>ones info
Thai CAE	complain about your numerous inaccurate <#RN><#CN> informations information #CN statements #RN and ask for a
Thai CAE	M> your publishing #M inaccurate and unfair <#CN> <mark>informations</mark> information #CN .
Thai FCE	I hope <#AGD>these this #AGD <#CN> informations information #CN <#AGV>are is #AGV helpful to
Frenc FCE	by infra-red link with my mobile phone to get <#CN> <mark>informations</mark> information #CN and electronic <#CN>mails mail
Frenc FCE	often go on <#MD> the #MD Internet to find <#CN> informations information #CN <#R>for when #R studying a <
Frenc FCE	ou <#R>needed asked #R <#MT> for #MT some <#CN>informations information #CN about me.
Germa FCE	I would like to get monthly <#CN> informations information #CN about any news concerning comp
Germa FCE	club I am looking for, please send me further <#CN>informations information #CN and don't forget to tell me th
Korea FCE	I will be glad, if this <#CN>informations information #CN <#AGV>help helps #AGV you to
Greek FCE	<pre><#DY>special especially<!--#DY--> in the area of <#CN>informations information<!--#CN-->.</pre>
Germa FCE	#M> and thank you in advance for your further <#CN>informations information #CN .
Itali CAE	<pre>Anyway<#MP> ,<!--#MP--> <#CN>informations information<!--#CN--> <#AGV>are is<!--#AGV--> given both</pre>
Thai FCE	<pre><#R>that in which<!--#R--> you asked me for some <#CN>informations information<!--#CN--> and some suggestions.</pre>
Itali FCE	er the rainbow", but <#RQ>many a lot of #RQ <#CN> informations information #CN <#AGV>were was #AGV wrong.
🛃 start	📓 Lancasterdraft - Micr 💱 OpenInsight for Work 🛐 Cambridge Internatio

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.