# Learner corpus linguistics in the EFL Classroom

**Peter Crosthwaite (prc34@cam.ac.uk)**

**Department of Theoretical and Applied Linguistics, University of Cambridge.**

## Bio:

The author has 8 years TESOL experience gained in South Korea and the U.K. He has worked as a university EFL lecturer, language teacher trainer and IELTS instructor/examiner, as well as director of studies for E.F. Language schools at Clare College, Cambridge. He holds an M.A. in TESOL from the University of London, an M.Phil. in applied linguistics from Cambridge University, and is now in the second year of doctoral studies in applied linguistics at Cambridge.

## Introduction:

Learner corpora, or searchable collections of the written or spoken production of language learners, have become increasingly vital in the field of applied linguistics. As advances in computational linguistics and natural language processing have allowed for the automatic or semi-automatic parsing, tagging and coding of large chunks of learner-produced text, there is increasing interest in the analysis of large-scale corpora of many millions of words, collected from a broad range of linguistic tasks, language backgrounds, and learner proficiencies.

Parsing and tagging are where text is analysed divided into parts of speech (tagging) and organised according to tokens (tagging) that determine grammatical or semantic structure. For example, if I parsed and tagged the sentence 'this is what corpus linguistics can do' (using the Stanford NLP parser, at http://nlp.stanford.edu:8080/parser/index.jsp), a tagger might produce the following output after analysing each word for POS (where elements such as DT = determiner, VB = verb, base form etc.) [i] :

```
This/DT is/VBZ what/WP corpus/JJ linguistics/NNS can/MD do/VB
```

The parser would then organise this sequence as below, allowing the dependencies and relationships of the words in the sentence to be explored [ii]:

```
(ROOT
  (S
    (NP (DT This))
    (VP (VBZ is)
      (SBAR
        (WHNP (WP what))
        (S
          (NP (JJ corpus) (NNS linguistics))
          (VP (MD can)
            (VP (VB do))))))
    (. .)))
```

As language teachers, we set our students a wide range of written and spoken tasks that, after completion, are forgotten about or deleted when the term ends.  In doing so, the opportunity to compare and actually describe what your students can (and cannot) do linguistically in *quantifiable* terms is wasted, and a valuable resource for shaping future curricula, lesson planning and assessments is lost.  This short ‘idea’ paper will outline some of the interesting work currently being undertaken by those working on learner corpora, how language educators can build their own learner corpora, and how the application of the findings from learner data can be used to build a detailed profile of the learners in your institution.

**What are learner corpora and what can teachers do with them?**

Learner data can come from a wide variety of sources, such as standardised tests, written assignments, student homework, internet chat room logs, recordings of oral data and many more.  By way of example, the Cambridge Learner Corpus (CLC) (www.cambridge.org/gb/elt/catalogue/subject/item2701617/Cambridge-Learner-Corpus) is a collection of 49 million words of learner data from 1993 to the present day, of which about half have been error-coded for a range of linguistic phenomena such as violations of syntactic structure, missing or incorrect use of determiners with noun phrases, inappropriate use of discourse markers, and many more besides (see Nicholls, 2003). The CLC contains data taken from learner production on a wide range of Cambridge ESOL exams, including IELTS, FCE, CAE, and CPE tests.  The data covers levels A1-

C2 of the Common European Framework for Languages (CEFR), which is the standard criterion-based framework of English proficiency levels for learning, teaching and assessment developed by the Council of Europe (2001a) (see Nicholls, 2003, Alexopolou, 2008, or Hawkins & Buttery, 2008, and Salamoura & Saville, 2010, for a detailed description of how this was achieved). This large collection of learner data is filtered according to different searchable criteria, such as the proficiency level of the learner, task type (essay, narrative etc.) or the first language of the learners themselves. The aim of the English Profile Programme behind the CLC is to build a profile of the inter-language of students at each proficiency level of English, and to characterize the associated patterns of errors that they make in their production. In other words, the project seeks to put in quantifiable linguistic terms what it really means to be at 'beginner', 'intermediate', or 'advanced' levels of English. Users of the corpus are able to see precisely what aspects of language are acquired at key stages of a learners' development, and the differences in how learners of different L1s structure their L2 English, i.e. to see transfer effects as and when they occur in real learner data. Among other projects, the CLC informed the development of the English Vocabulary Profile which shows the kind of words and phrases to be found at each proficiency level (A1-C2) of the CEFR. This is resource can be accessed online at http://www.englishprofile.org/index.php?option=com_forme&Itemid=107. The organisers of the English Profile Project still accept submissions from institutions who wish to contribute their data to the project at http://www.englishprofile.org/index.php?option=com_forme&Itemid=77.

**Research Example:**

As part of background (unpublished) studies into my Ph.D. research, I used the CLC to investigate how coherent reference to discourse entities was achieved in Chinese and Korean English learners' L2 narrative discourse. By reference, I mean any linguistic expression that is used to introduce or maintain reference to a character in a discourse text, including expressions with indefinite articles ('a boy'), definite articles ('the boy'), pronouns ('he') or otherwise. It is suggested that L1 learners share certain universal tendencies in the production of referring expressions in discourse (Hickmann et. al. 1996, Hickmann & Hendriks, 1999) despite large typological differences between languages in how these expressions are realised within individual languages. However, these differences mean that L2 learners may encounter L1 transfer or L2 learnability issues when attempting to produce referring expressions that follow the norms of the target language, even if the intentional/relational structure of the discourse is controlled for by the question posed in the task description. These issues may be more apparent in adult L2 learners, who already have fully-developed strategies for maintaining the coherence of referring expressions in their L1, but may have passed the 'critical period' for language acquisition that allows them to re-align their linguistic strategies to manage reference in a way that ensures coherence in the target language.

I searched the Cambridge Learner Corpus to look at L2 English narrative texts from Korean and Chinese English learners over four proficiency levels (B1-C2 of the CEFR) for their use of referring expressions. I did this to ascertain which and when L1 transfer effects occurred and if and when these learners' use of referring expressions eventually became target-like. As the CLC has already been error coded, I analysed the frequency and type of errors that were produced by the Chinese and Korean learners at each proficiency level and found that more errors of reference were found at B1 than C2 level in both languages, and that Koreans on the whole produced more errors when producing referring expressions than their Chinese counterparts at B1 levels in terms of missing anaphor (MA)($f$=5.508, $p$<0.05) and at C2 level in terms of incorrect determiners (RD)($f$=8.608, $p$<0.05), suggesting a higher degree of L1 transfer in their L2 production than their Chinese counterparts. In addition, the kind of error made by both groups of learners followed a developmental path from frequent missing determiners (where articles such as 'a' and 'the' were missing) in texts at B1 level, to choosing the wrong kind of determiner (e.g. using 'those' instead of 'the' when referring to 'the boy') at higher (C2) levels, as described in the charts below:

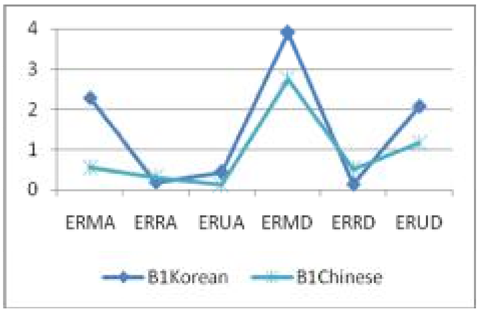

*Fig 1.* Errors at B1 level as a percentage of the total referring expressions made[1]

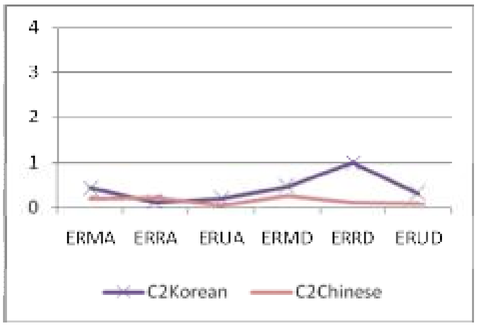

*Fig 2.* Errors at C2 level as a percentage of the total referring expressions made

By analysing the learner data gathered in the corpus in this way, I was able to characterize in *quantifiable* terms the errors these learners had made at each proficiency level, which allowed me to make judgements about the influence of L1 transfer effects on L2 production between these two source languages.

More recently, a new project between the Department of Theoretical and Applied Linguistics at Cambridge University and Education First Language Schools, titled 'the EF-Cambridge Corpus of Learner English' (Post, Alexopoulou, Geertzen & Korhonen, 2012) uses EF's online teaching and learning platform 'Englishtown'. In this project, a range of learner data from homework assignments, chat room logs, assessments and speech data will be used for syntactic, semantic, discourse and phonological and phonetic analysis, with the potential for a vast amount of *continuously updated* learner data for researchers to analyse in the near future.

[1] Key - ERMA = Missing Anaphor (ex: __ went out), ERRA = Replace Anaphor (ex: 'he' instead of 'she' for female referents), ERUA = Unnecessary Anaphor (ex: 'he she went out') ERMD = Missing Determiner (ex: 'boy went to school'), ERRD = Replace Determiner (Ex: 'a boys went out') ERUD = Unnecessary Determiner (ex: 'the lions are wild animals' in the generic sense).

Clearly the potential benefits of being able to analyse large collections of learner data are massive for research into language acquisition and learner performance, and this research will feed innovation in curriculum development and assessment of language proficiency.

### How learner corpora can be built:

While large projects such as the English Profile Project are seeking to answer the fundamental questions of language learning, learner corpora do not have to be so large to be of use to educators in their own contexts. Although ‘bigger’ may be ‘better’ in terms of the generalisability of the statistical analysis of the data (where many millions of words are required), in certain EFL contexts, it is the inter-individual data that may be of more importance, and as such smaller amounts of data (in the region of tens of thousands of words) may be sufficient. However, as more schools introduce computer labs with CALL, tablet PCs and IPads to enhance the learning of their students, an exponential amount of data is being submitted to language teachers in an electronic format. As time goes by, improvements in spoken word recognition are easing the process of converting spoken data into electronic text that can be stored for the purpose of building personal collections of learner data into searchable learner corpora. However method is used to convert the learner data, once this data has been made available as searchable text, using the corpus to find the data you need should then be possible.

### Choosing the categories:

When building a corpus, it is advisable to first consider the categories that the data will be divided into, and assign headers to the data so that it can be searchable according to defined criteria. For example, in many teaching contexts, learners are divided into ‘beginner’, ‘intermediate’ or ‘advanced’ proficiencies, or data can also be separated by learner age, whether they are child or adult learners, learners of different L1 backgrounds, different genders etc. Another strategy is to observe the longitudinal development of a single or group of learners’ progress in a particular class over time, by adding learner I.D. and date information to the headers of the data. For English for Specific Purposes (ESP), task type (narrative, discursive essay, newspaper report, debate, ‘map’ task etc.) is a useful category to observe in order to see where students have had little experience or where they face particular difficulties. Another useful idea would be to label data as 'pre-' or 'post-' lesson data, in order to see if a particular lesson (or technique used in that lesson) caused any real quantifiable change in your learners' production of the target form, assuming that the teacher's input was also going to be added to the corpus in this case.

This kind of data is directly useful for future teacher training and / or lesson planning, and can also allow the teacher to take on the role of a researcher into the effects of pedagogy on linguistic production. However the data is divided may also influence how the data will be stored, i.e. whether the corpus is a single large collection across the institution, or divided into a number of sub-corpora that can be used by individuals working on different projects within the department. Not all of the variables suggested here need to be included into the metadata (e.g. L1 data for monolingual contexts, what to include is ultimately a matter of how your institution wishes to use the data.

**Corpus Coding:**

The next thing to consider when building a corpus of learner data is how the coding of the data will be handled. This really is where the corpus administrator must ask the question "what do I want to know about what my learners are producing?" Commonly, corpus coding can take the form of syntactic analyses (or *parsing*) of the syntactic structure of learner production, such as if they are correctly using SVO structure in their English sentences, correctly using clause subordination, or whether they are missing determiners from their noun phrases. In addition, language *functions* that a particular learner performs during their production, such as the ability to offer an opinion, decline a request, agree with a statement etc. can be coded for frequency and acceptability. While error coding can be useful as demonstrated above, often treating the structure of the learners' interlanguage as an basic form of the target language itself may be a more useful point of comparison than simply seeking for 'errors' *per se*. This is because looking at what they 'are' doing rather than talking about what they are 'not' doing can be a more fruitful investigation of a learner's capabilities. In my investigation into referential coherence, I utilised the coding system of Hickmann, Hendriks, Roland &

Liang (1994) for referential expressions, and applied them to transcriptions of spoken narrative data, as with the example below from Korean data:

```
*P:       학교에서 한 친구가 농구공을 가지고 +.                (the actual learner production)

 %rom:    hakgyoeseo han chinguga nonggugongeul gajigo       (the Romanisation of this production)

%tra:     school at one friend sub basketball obj have and   (the word-for-word translation)

 %eng:    a friend had a basetball at school and             (the English translations)

 %cod:    $REL:FOREGROUND:INTRODUCTION
$R1:FM:NUMNOM:SUBMARK:POS:BOY1:PREV:one+friend
```

The final line (%cod:) contains the coding system used, with $REL: showing the relational token of each sentence (in this case, this sentence is the introduction to the narrative, and is in the foreground) and $R1: displaying the referential data of the reference to the 'friend' character, with FM being first-mention, NUMNOM being numeral+nominal (one + friend), SUBMARK showing that the noun is marked as a subject in Korean (with the '가' suffix), POS showing the semantic role of the noun (that he is the possessor of the ball), BOY1 as the actual character, PREV to show that the reference was pre-verbal (with implications for differences in word order between source and target languages), followed by the actual NP construction. In this way, I can analyse the coherence of the references made in the discourse over an individual text, and make generalisations about references used in the cohort of texts from each proficiency level. By thinking carefully about what needs to be improved in a learner's performance, coding of the learner data should reflect the questions that need to be answered about their production.

Hand-coding data can be time-consuming, although modern natural language processing applications can do most of the heavy work automatically - in terms of breaking down language into its key elements, such as part of speech, syntactic structure, semantic roles etc. Part of speech 'taggers' automatically scan text and tag it as verbs, nouns etc, such as Treetagger (Schmidt, 1994), or CLAWS (Garside, 1987). Other online resources such as Verbnet (Kipper-Schuler 2006) can be used to help corpus builders tag semantic/thematic roles (such as agent/patient/manner etc.) onto learner data in order to assist with the analysis of the argument structure of verbs (i.e. whether the learners' are using transitivity correctly), and DEXTER (Garretson, 2005) allows for manual coding and annotation of written and spoken data for coding and analysis. ANTCONC (Antony, 2006) allows for condordances, word lists, collocates and keyword functions to be drawn from data for analysis.

The above links are designed for native language and therefore may have lower accuracy rates in learner data, yet are good starting point for an investigation into learner data. Despite the sometimes time-consuming nature of coding a corpus, by investing the time into coding learner data, one may build a permanent resource that will produce quantifiable 'hard evidence' of your learners' abilities. Coding the data is something that even learners themselves may be willing to assist with.

### Using the Corpora

As explained above, the uses of learner corpora are really only restricted by how one intends to code them. When coded however, it is then possible for users of a learner corpus to quantify linguistic phenomena within learner data that can be used to compare the production and performance of individuals and groups of learners, and these comparisons and statistics can be used to drive lesson planning, assessment and long-term curriculum development. By way of example, Flowerdew (2001) highlights how concordances (or frequency lists of co-occuring words) can be used to look at patterns of collocation with learner data with applications for the teaching of vocabulary. Granger (2002) summarizes a large body of ELT related research using learner corpora that includes dictionary building and cross-linguistic error analysis. Seidlhofer (2002) uses learner corpora to drive pedagogy through the use of actual texts from learner corpora in class, in a student-led linguistic analysis task. A large collection of learner corpora-driven research can be found at http://www.uclouvain.be/en-cecl-lcBiblio.html which may be interest when deciding how learner corpora can aid in building a profile of the learners in your context.

### Closing comments:

Data from learner corpora is finally starting to bring the work applied linguists and classroom professionals in the same direction: to explore how languages are acquired, when they are acquired, and what is actually acquired at each stage of the learning process. Rather than wasting such a golden opportunity to learn about the learners in your educational context, we should exploit the rich and varied benefits that statistical analyses of learner production data can bring about, and make the call for teachers and researchers alike to utilize the technology (that is now increasingly available) to unlock the linguistic profiles of language learners worldwide.

## Acknowledgments

I would like to thank Akira Murakami of the Department of Theoretical and Applied Linguistics, University of Cambridge, for his invaluable comments on the NLP tools available for learner corpora. The Cambridge Learner Corpus was developed jointly by Cambridge ESOL and Cambridge University Press, and forms part of the Cambridge English Corpus

## References:

Alexopoulou, T. (2008). Building new corpora for English Profile. *Research Notes,* (33), 15-19.

Council of Europe (2001a). Common European Framework of Reference for Languages: Learning, Teaching, Assessment. Cambridge: Cambridge University Press.

Flowerdew, J. (2001). The exploitation of small learner corpora in EAP materials design. In M. Ghadessy, R. Roseberry & A. Henry (Eds.), *Small corpus studies and ELT* (pp. 264-79). Amsterdam: John Benjamins.

Granger, S. (2002). A bird's-eye view of computer learner corpus research. In S. Granger, J. Hung, S. Petch-Tyson & J. Hulstijn (Eds.), *Computer learner corpora, second language acquisition and foreign language teaching* (pp. 3-33). Amsterdam & Philadelphia: John Benjamins.

Hawkins, J., & Buttery, P. (2008). *Using learner language from corpora to profile levels of proficiency - insights from the English profile programme*. Paper presented at the Third ALTE Conference 2008, Cambridge (UK).

Hickmann, M., Hendriks, H., Roland, F., & Liang, J. (1994). *The development of reference to person, time, and space in discourse: a coding manual.* Max Planck Institute for Psycholinguistics.

Hickmann, M., Hendriks, H., Roland, F., & Liang, J. (1996). The marking of new information in children's narratives: a comparison of English, French, German and Mandarin Chinese*. *Journal of child language, 23*(03), 591–619.

Hickmann, M., & Hendriks, H. (1999). Cohesion and anaphora in children's narratives: a comparison of English, French, German, and Mandarin Chinese. *Journal of Child Language*, *26*(02), 419–452.

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, UCREL technical paper number 16*, UCREL, Lancaster University.

Post, B., Alexopoulou, D., Geertzen, G. Korhonen, A. (2012). The EF Cambridge Corpus of Learner English: L2 speech data. Conference paper presented at *IVACS Symposium: Speech is Where It's At*, University of Cambridge.

Salamoura, A. & Saville, N. (2010) Exemplifying the CEFR: Criterial features of written learner English from the English Profile Programme. In Bartning, I, Maisa, M & Vedder, I (Eds.) *Communicative proficiency and linguistic development: Intersections between SLA and language testing research*. EuroSLA Monographs Series (1), 101-132.

Seidlhofer, B. (2002). Pedagogy and local learner corpora: Working with learning-driven data. *Computer learner corpora, second language acquisition and foreign language teaching*, 213–234.

**Web Links:**

Antony, L. (2006), ANTCONC, http://www.antlab.sci.waseda.ac.jp/software.html

Garretson, G. (2005), DEXTER – Tools for analyzing language data, http://www.dextercoder.org/about.html)

Garside, R. (1987) CLAWS part-of-speech tagger for English, http://ucrel.lancs.ac.uk/claws/

Kipper-Schuler, K. (2006), Verbnet - A Class-Based Verb Lexicon, http://verbs.colorado.edu/verb-index/index.php

Schmidt, H.(1994) TreeTagger - a language independent part-of-speech tagger, http://www.ims.uni-stuttgart.de/projekte/corplex/TreeTagger/

[i] A list of the abbreviations for the Stanford tagger are found at http://www.computing.dcu.ie/~acahill/tagset.html

[ii] The definitions of the abbreviations for the Stanford parser are found at http://nlp.stanford.edu/software/dependencies_manual.pdf