Investigating the Cambridge Learner Corpus

by Annette Capel, co-author of the Objective series

From Objective KET to Objective Proficiency

In writing these preparation courses for the Cambridge ESOL examinations, Wendy Sharp and I have been fortunate to have access to the Cambridge Learner Corpus, a vast electronic collection of examination scripts. This unique resource has given us detailed evidence and insights into what candidates at each Cambridge level can and can't do. No other course books for the Cambridge examinations are 'corpus-informed' in this way, making the Objective series rather special.

As a Cambridge examiner, I must have marked many, many thousands of exam scripts at different levels over the last fourteen years and through this, I have been able to develop some awareness of candidate competence worldwide, particularly at B1, B2 and C1 levels, since these are the exams I have marked regularly (PET, FCE and CAE). However, working with the Learner Corpus has been a quantum leap for me, in terms of the comprehensive picture offered.

What is the Cambridge Learner Corpus?

The corpus, jointly funded by Cambridge University Press and Cambridge ESOL, is part of a much larger text collection called the Cambridge International Corpus, which we also have access to as writers. The Learner Corpus has been growing steadily since 1993 and at the time of writing contains over 18 million words of candidates' writing. All levels and exams are represented, from KET and PET up to CPE, as well as BEC, CELS and IELTS scripts. It is possible to search the corpus at single word or phrase level, and to obtain exact frequency information. Moreover, each candidate's first language is recorded, enabling the user to filter the corpus according to very specific requirements. For example, I recently gave a talk at IATEFL Slovenia, for which I was able to extract data purely from Slovene speakers for a defined range of exams.

In the first instance, the Learner Corpus provides concordance lines, with the search word highlighted in red and running down the middle of the computer screen. These lines can then be sorted in various ways, such as to the left or right of the word, which makes it easier to spot common collocates and grammatical patterns. Sometimes, the limited context provided within a concordance line is insufficient, but it is always possible to call up more text: a complete sentence, a paragraph or indeed a whole script.

Coded data

What makes the Learner Corpus doubly useful for someone like me is the fact that a substantial amount of the data is tagged, with a very detailed system of error codes in place. This allows for close scrutiny of anything from spelling errors to problems of word order, from noun agreement to tense errors, and from confused words to inconsistent register, all available instantly and abundantly, because it is possible to search the Learner Corpus on a single error code.

It may come as no surprise to anyone preparing students for the Cambridge exams that the word ‘beautiful’ is one of the words candidates most frequently mis-spell. This word seems to cause problems throughout the Cambridge levels, right up to C2 Proficiency, and it was certainly one of the top spelling errors we noticed when we researched the KET data last year, during the preparation of our new course Objective KET (to be published early 2005). The corpus actually gave us a ‘hitlist’ of other problematic words containing two or three vowels together, which we chose to focus on in one of the ‘Spelling spot’ sections of the Student’s Book.

Accuracy, appropriacy and range

Clearly, the error codes have been of great help to us as course book writers, allowing us to prioritise obvious problem areas and decide on final coverage. However, for me, accuracy is only one side of the coin; corpus information on appropriacy and language range is equally welcome, for it demonstrates what candidates actually **can** do at each level. Incidentally, the corpus data supports the Common European Framework ‘Can do’ statements for Writing at each level very convincingly.

Appropriacy is an area of language expertise that students gradually get to grips with and it only really begins to be an examining issue from B2 level onwards. At the top end it is an essential requirement for success and so, when we were writing Objective Proficiency, we inevitably focused on style and register, and were able to base many useful exercises and tasks on the candidate data itself.

When browsing the Learner Corpus, it is often extremely heartening to see what candidates have been able to produce under exam pressure. The range of structures and vocabulary exemplified by the corpus is impressive and, because language ambition is recognised and rewarded within the marking criteria of the Cambridge exams, we have consciously tried to develop good productive range in all the Objective titles we have written. Our Writing Folders deal with this aspect systematically, and we also encourage students to extend their range and experiment in the classroom in preparing them for a Cambridge Speaking test.

The Objective series – state-of-the-art exam preparation

As you can see, our Objective titles are underpinned by precise evidence, in the shape of the written performance of recent Cambridge ESOL candidates. This is part of our recipe for success, combined with lively topics in motivating short units. If you haven’t yet joined the Objective club, why not try out one of our courses for yourself? Give your students the very best chance of exam success!

Annette Capel 22.10.04