Learner corpus research: some problems, some questions, and some possible answers

Hilary Nesi
Coventry University

This talk will explore issues associated with the notion of the learner corpus, illustrated with references to my own experience as a language teacher, language learner, researcher and corpus designer.

First of all, it will ask how we should define the 'learners' who produce learner corpus content. There seem to be three basic ways of deciding this - by self-identification, by mother-tongue status, or according to their presence in a language learning class. The first two definitions are a bit problematic, as in some respects we can all self-identify as learners, and most people in the world have mixed proficiencies in more than one language. Many people report that they are happier using one language at home and in their own cultural contexts, and another language when communicating their academic or professional expertise, especially if they have acquired their expertise in the medium of this other language. The 'native speaker' designation is increasingly rejected by educationalists and journal editors because it implies superior communication skills in the mother tongue, something we know is by no means guaranteed. On the other hand, if learners are only learners when performing in the language learning class, the only texts that can be included in a learner corpus are those produced for the purposes of language learning or assessment. There is a danger that such texts will be coloured by the demands of the language learning syllabus, with certain linguistic features included only for the purposes of display.

In light of this, we also have to decide on appropriate methods of learner corpus analysis. Most approaches, beyond identifying typical structural errors, require some comparison with texts produced by 'non-learners' – probably people who use the language as their mother tongue, people considered expert speakers or writers, or both. Corpus compilers know that it can be rather difficult to identify who is and who is not a native speaker, especially in studies involving large numbers of texts, perhaps produced by multiple authors. Moreover, any comparison between texts produced in different situational contexts automatically introduce extra variables that have nothing to do with language learning status: whether we compare texts produced by experts with those produced by novices; local texts with those produced for international audiences; or texts produced in the language classroom with those produced for any genuine academic, professional or social purpose.

The best thing seems to be to address these problems full on, acknowledging the inherent difficulties in learner corpus research and making allowances for them when designing corpora and drawing our conclusions. I hope the talk will be thought-provoking, and give rise to some lively discussion.