A marriage of two minds? Learner translation corpora in learner corpus research

Silvia Bernardini
Università di Bologna

Learner corpus research (LCR) is understood, quite naturally, as the adoption of corpus linguistics techniques in the study of language learning and acquisition, in other words for describing and modelling non-native or second language varieties through the investigation of learners' production. The field of corpus-based translation studies, which has come to the fore and developed in parallel to LCR, also aims to conceptualize and investigate what is purported to be a separate language variety, namely translated language. The two fields thus have several points of contact, that have recently led to a partial alignment of interests and priorities. As a result, several learner translation corpora (LTCs) have seen the light, to which well-established practices from LCR (such as error annotation) are also applied.

In this talk, I will first of all discuss the ways in which LTCs can be of interest to the field of LCR and language pedagogy at large. First, pedagogic translation has consistently been practiced in the language classroom, and has even been rehabilitated in recent years. Second, translation data provide direct first language equivalents for produced second/target language segments, which may complement datasets resulting from freer production tasks. Third, and more importantly, bringing the two research frameworks together allows one to pursue the fascinating, and very ambitious, goal of understanding similarities and differences between L2 and translated production seen as instances of constrained communication in language contact situations.

Somewhat provocatively, I will also point out several important methodological and theoretical issues. Indeed, notwithstanding the potential advantages of this alignment, the nature of the data is such that one may wonder whether it is legitimate to include current LTCs fully among learner corpora, or even to consider them corpora at all. To illustrate my point, I will refer to an exploratory attempt at combining translation and essay writing data in the exploration of English topic-neutral, high-frequency collocations. The datasets used are not closely comparable in terms of topic and register, since they contain texts assembled from previous coursework and examinations: a suboptimal, yet rather common condition applying to non-experimental settings in which translated and non-native language varieties are compared.

Rather than provide answers to such complex questions, I hope to stimulate discussion on how we can make sense of learner copora and LTCs, and of complex datasets representing multiple instances of learner production in general, while remaining true to the methodological and theoretical assumptions informing corpus linguistics.