As predicted by Mona Baker almost two decades ago (Baker 1993), the rise of corpus linguistics has had a great impact on translation studies over the past years. Bilingual comparable corpora have been used in translation training (e.g. Kübler 2011) and to explore functional equivalents between languages (e.g. Tognini Bonelli & Manca 2004). Comparable corpora of translated and non-translated texts have been used to study what sets translated language apart (e.g. Laviosa 1998, Olohan & Baker 2000, Tirkkonen-Condit 2004, Puurtinen 2004, Frankenberg-Garcia 2008). And parallel corpora have been used in various studies comparing source-texts and translations (e.g. Frankenberg-Garcia 2005, Johansson 2007, Frankenberg-Garcia 2009, Pérez Blanco 2009, Lefer 2010, Bernardini 2010, Saldanha 2011). However, most existing corpus-based analyses focusing on the contrasts between source texts and translations are either purely lexical or are constrained by sentence boundaries. Probably the main reason underlying this limitation is the fact that, in parallel corpora, texts are usually segmented at the level of the sentence due to the relative ease with which sentence boundaries can be identified automatically. The alignment of source texts and translations is then usually carried out such that whenever there is not a one-to-one correspondence between source-text and translation segments, they are aligned either on a one-to-many or on a many-to-one basis, blurring out the details of what happens when source-text sentences are not preserved in translation. With the focus of this 6th Corpus Linguistics conference being on discourse, the study of which transcends the level of the sentence, this paper aims to explore shifts in sentence boundaries that occur in the process of translation.

The unique, manually post-edited alignment of COMPARA - a three-million-word parallel corpus of English and Portuguese fiction (Frankenberg-Garcia & Santos 2003) - enables one to retrieve automatically all sentences that were joined together and all sentences that were split apart in translation. Using this free online corpus, the present exploratory study seeks to answer questions such as (1) To what extent do translators preserve sentence boundaries? (2) What is more common in translation: sentence-splitting or sentence-joining? (3) Do English and Portuguese language translators differ in terms of preserving sentence boundaries? (4) Could author or translator style bear an influence on the extent to which sentence boundaries are preserved or not? (5) What kind of sentences do translators split and what kind sentences do they join together? It is believed that the findings observed can have implications not only for translator training, but also for bilingual text alignment and the development of machine translation and translation memory systems.


