and Ru-En language pairs and the multilingual track. For the other language pairs (En-De, Ro-En, Et-En and Si-En), it shares first place with (Fomicheva et al. 2020a), whose results are not statistically different from ours. Therefore, TransQuest is the new state-of-the-art in sentence-level QE.

9.4 Error analysis

In an attempt to better understand the performance and limitations of *TransQuest* we carried out an error analysis on the results obtained on Romanian - English and Sinhala - English. The availability of native speakers determined the choice of language pairs we analysed to perform this analysis. We focused on cases where the difference between the predicted and expected scores was the greatest. This included instances where the predicted score was underestimated and overestimated.

Analysis of the results does not reveal obvious patterns. The largest number of errors seem to be caused by named entities in the source sentences. In some cases, these entities are mishandled during the translation. The resulting sentences are usually syntactically correct but semantically odd. Typical examples are *RO: În urmă explorărilor Căpitanului James Cook, Australia și Noua Zeelandă au devenit ținte ale colonialismului britanic. (As a result of Captain James Cook's explorations, Australia and New Zealand have become the targets of British colonialism.) - EN: Captain James Cook, Australia and New Zealand have finally become the targets of British colonialism.* (expected: -1.2360, predicted: 0.2560) and *RO: O altă problemă importantă cu care trupele Antantei au fost obligate să se*

confrunte a fost malaria. (Another important problem that the Triple Entente troops had to face was malaria.) - EN: *Another important problem that Antarctic troops had to face was malaria.* (expected: 0.2813, predicted: -0.9050). However, it is debatable whether the expected scores for these two pairs should be so different. Both of them have apparent problems and cannot be clearly understood without reading the source. For this reason, we would expect that both of them have low scores. Instances such as this also occur in the training data. As a result of this, it may be that *TransQuest* learns contradictory information, which in turn leads to errors at the testing stage.

A large number of problems are caused by incomplete source sentences or input sentences with noise. For example the pair *RO: thumbright250pxDrapelul cu fâșiile în poziție verticală (The flag with strips in upright position) - EN: ghtghtness 250pxDrapel with strips in upright position* has an expected score of 0.0595, but our method predicts -0.9786. Given that only *ghtghtness 250pxDrapel* is wrong in the translation, the predicted score is far too low. In an attempt to see how much this noise influences the result, we ran the system with the pair *RO: Drapelul cu fâșiile în poziție verticală - EN: Drapel with strips in upright position.* The prediction is 0.42132, which is more in line with our expectations, given that one of the words is not translated.

Similar to Ro-En, in Si-En, most problems seem to be caused by the presence of named entities in the source sentences. For example, in the English translation: *But the disguised Shiv will help them securely establish the statue.* (expected: 1.3618, predicted: -0.008), the correct English translation would be