Technical Problems of BLEU

BLEU scores are not comparable:

- across languages.
- on different test sets.
- with different number of reference translations.
- with different implementations of the evaluation tool.
- There are different definitions of “reference length”:

  Papineni et al. (2002) not specific. One can choose the shortest, longest, average, closest (the smaller or the larger!).
- Very sensitive to tokenization:

  Beware esp. of malformed tokenization of Czech by foreign tools.

⇒ Use a fixed implementation, e.g. sacreBLEU (Post, 2018).