Ultimate Goal of Traditional SMT

Find **minimum translation units** (MTUs) $\sim$ graph partitions:

- such that they are frequent across many sentence pairs.
- without imposing (too hard) constraints on reordering.
- (ideally in an unsupervised fashion, no reliance on linguistics).

Available data: Word co-occurrence statistics:

- In large monolingual data (usually up to 10^9 words).
- In smaller parallel data (up to 10^7 words per language).
- Optional automatic rich linguistic annotation.