Sub-Words to Reduce Vocabulary Size

- SMT struggles with productive morphology (>1M wordforms).

  nejneobhodpodařovávatelnějšími, Donaudampfschifffahrtsgesellschaftskapitän

- NMT can handle only 30–80k dictionaries.

⇒ Resort to sub-word units.

Orig	český politik svezl migranty
Syllables	čes ký ⊔ po li tik ⊔ sve zl ⊔ mig ran ty
Morphemes	česk ý ⊔ politik ⊔ s vez l ⊔ migrant y
Char Pairs	če sk ý ⊔ po li ti k ⊔ sv ez l ⊔ mi gr an ty
Chars	č e s k ý ⊔ p o l i t i k ⊔ s v e z l ⊔ m i g r a n t y
BPE 30k	český politik s@@ vez@@ l mi@@ granty

BPE (Byte-Pair Encoding, (Sennrich et al., 2016)) or Google's wordpieces (Wu et al., 2016) and Tensor2Tensor's SubwordTextEncoder use n most common substrings (incl. frequent words).