Tokenization Analysis: WordPiece

Token Occurrences by Length Quantile

Y-axis: Token Occurrences
X-axis: Vocabulary Size

Quantiles:
- Q25
- Q40
- Q50
- Q60
- Q75
- Q100
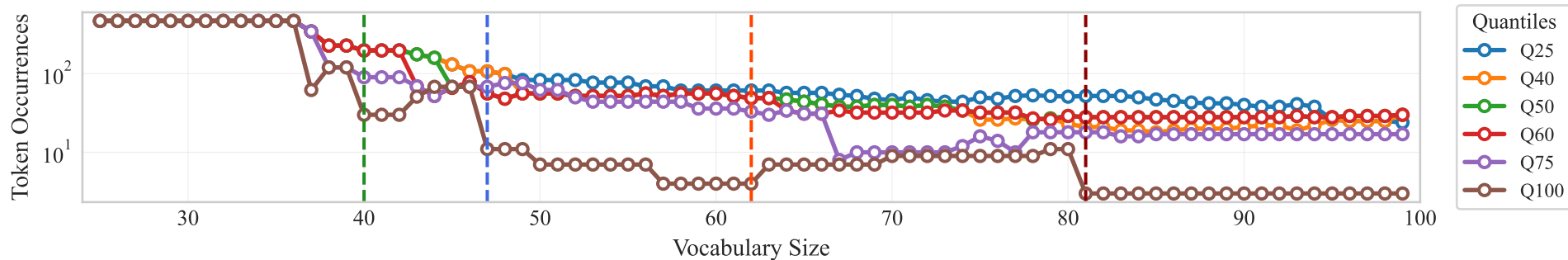
Token Frequency Normalized by Length

Y-axis: Occurrences / Length
X-axis: Vocabulary Size

Labels: AB, SY, ZAB, TAB, SYU, VE, ST, ZABSY, TA, ESY, ABSY, ABSYU, ZA, VE, SYV
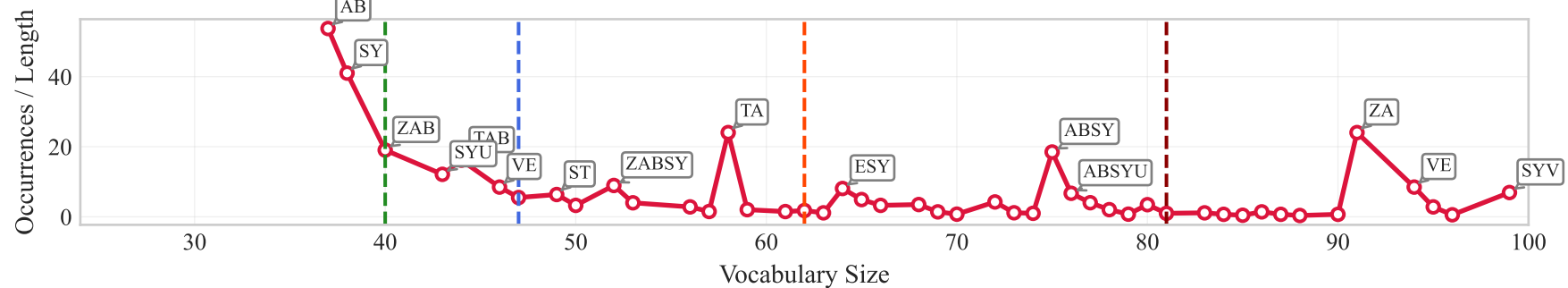
Shannon Entropy of Token Distribution

Y-axis: Entropy (bits)
X-axis: Vocabulary Size

Reference Points:
- Vs=40
- Vs=47
- Vs=62
- Vs=81