Byte Pair Encoding: Learning Subwords
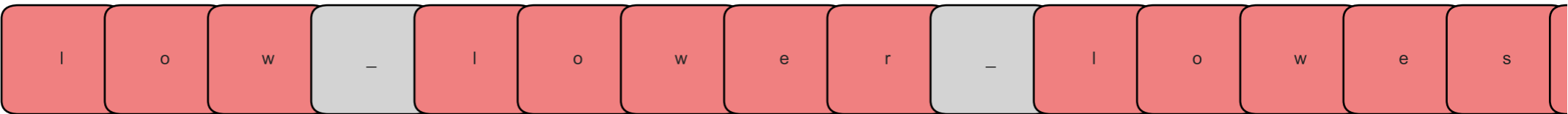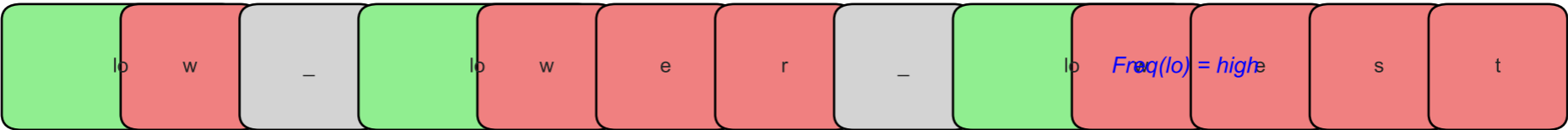
Corpus: "low lower lowest"
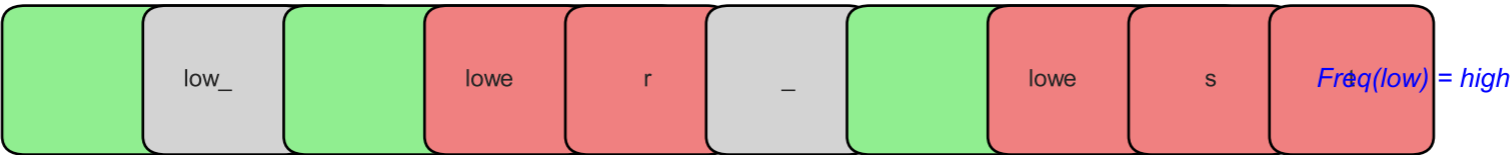
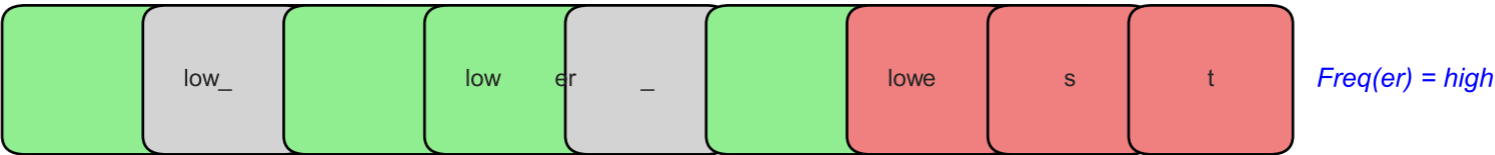**Initial**

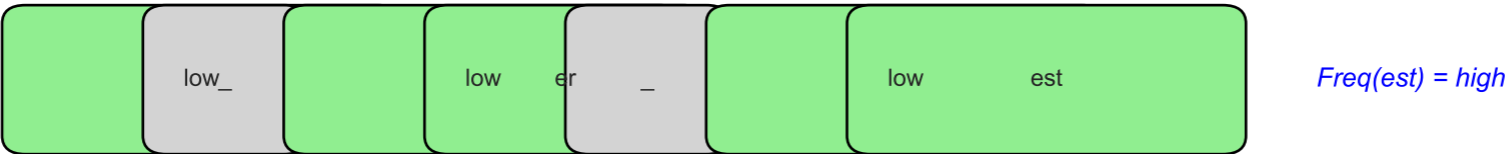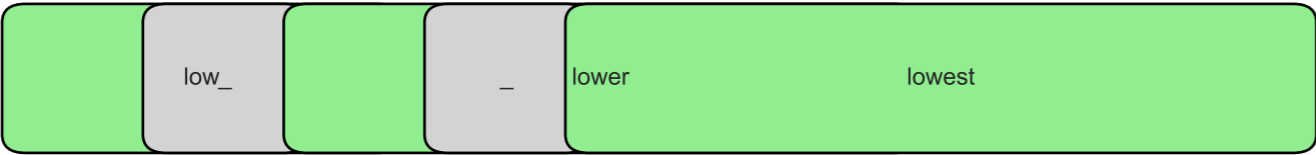| l | o | w | _ | l | o | w | e | r | _ | l | o | w | e | s | t |

Merge "lo"

| lo | w | _ | lo | w | e | r | _ | lo | w | e | s | t |

Freq(lo) = high

Merge "low"

| low_ | | lowe | r | _ | lowe | s | t |

Freq(low) = high

BPE discovers:
- "low" = root
- "er" = comparative
- "est" = superlative

Merge "er"

| low_ | low | er | _ | lowe | s | t |

Freq(er) = high

Merge "est"

| low_ | low | er | _ | low | est |

Freq(est) = high

Final
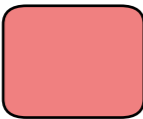
| low_ | | _ | lower | lowest |

| Character | Subword | Space |