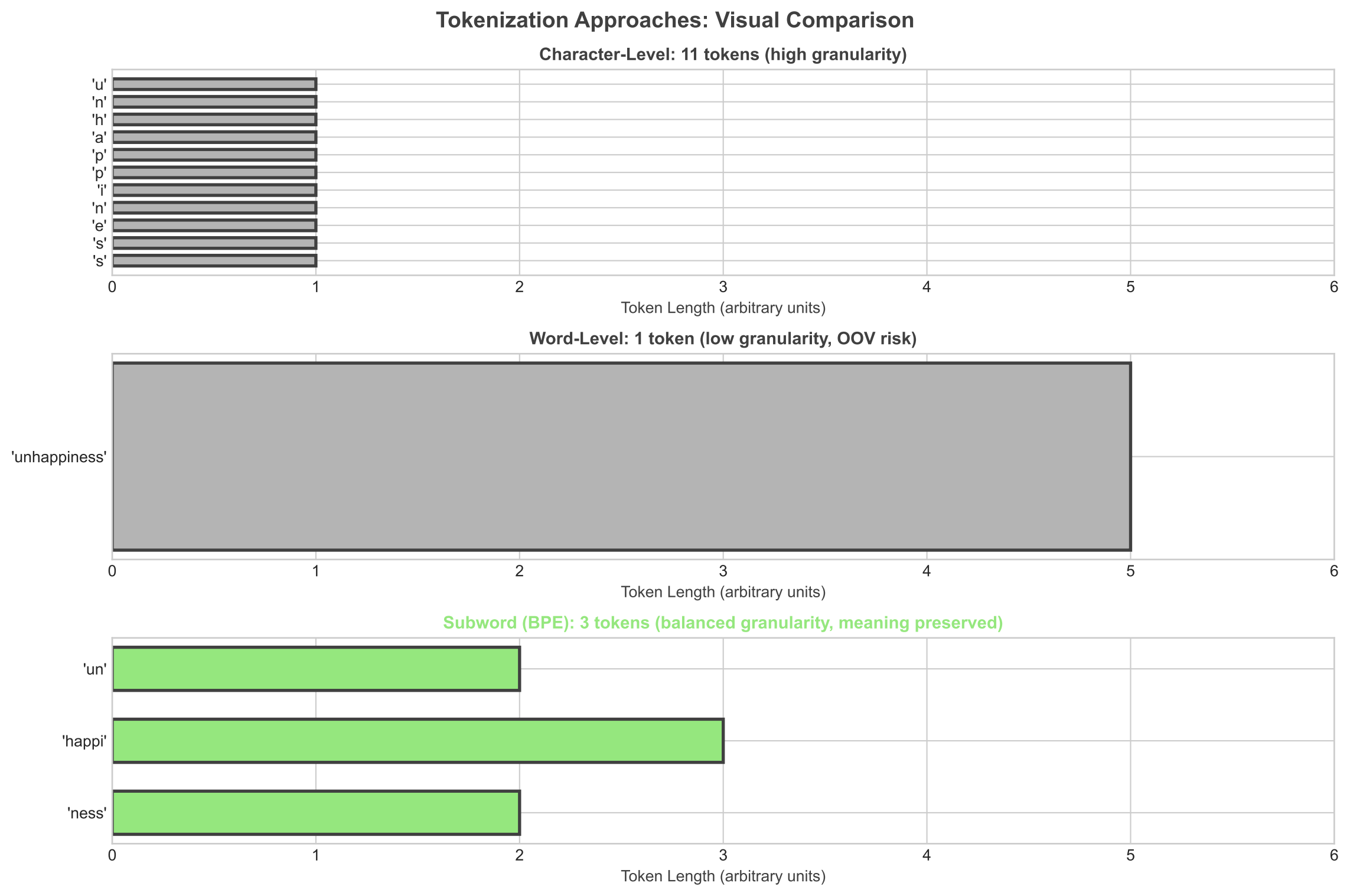
Tokenization Approaches: Visual Comparison
Character-Level: 11 tokens (high granularity)
'u'
'n'
'h'
'a'
'p'
'p'
'i'
'n'
'e'
's'
's'
0
1
2
3
4
5
6
Token Length (arbitrary units)
Word-Level: 1 token (low granularity, OOV risk)
'unhappiness'
0
1
2
3
4
5
6
Token Length (arbitrary units)
Subword (BPE): 3 tokens (balanced granularity, meaning preserved)
'un'
'happi'
'ness'
0
1
2
3
4
5
6
Token Length (arbitrary units)