## **Data exploration / linguistic indicators**

**Results**

**(1) Frequency of ngrams for avoidance / non-avoidance and different strategies (using FightingWords)**

**1.1 Avoidance / non-avoidance**

* Non-avoidance:
  + 10 most frequent bigrams in non-avoidance answers: ['you know', 'it was', 'lot of', 'if you', 'you re', 'you have', 'little bit', 'and you', 'going to', 'and then']
* Avoidance
  + 10 most frequent bigrams in avoidance answers: ['of the', 'the hon', 'my hon', 'hon gentleman', 'that we', 'hon friend', 'that is', 'will be', 'has been', 'per cent']

→ all with significant z-scores

**1.2 ‘Fight’ vs. ‘Flight’**

* Fight:
  + 10 most frequent bigrams in ‘Fight’ answers: ['the hon', 'hon gentleman', 'hon lady', 'do not', 'we will', 'that have', 'we have', 'the government', 'that he', 'that we']

→ only first three with significant z-score

* Avoidance
  + 10 most frequent bigrams in ‘Flight’ answers: ['you know', 'on the', 'at the', 'hon friend', 'with the', 'have to', 'of course', 'is the', 'there are', 'is very']

→ none with significant z-score

### **(2) Levenshtein distance between question and answer**

Threshold refers to the length difference between question and answer. When one of the two is much longer than the other, the Levenshtein distance will also increase.

**2.1 Avoidance vs. non-avoidance**

* Average Levenshtein distance for non-avoidance Q-A pairs:
  + Threshold 20: 102.31
    - 10% percentile: 22.400000000000002
    - 90% percentile: 205.39999999999998
  + No threshold: 326.47
    - 10% percentile: 91.0
    - 90% percentile: 602.0
* Average Levenshtein distance for avoidance Q-A pairs:
  + With thresholds 20, 50: 280.21
    - 10% percentile: 38.400000000000006
    - 90% percentile: 428.6
  + No thresholds: 459.34
    - 10% percentile: 136.1
    - 10% percentile: 599.2

→ non-avoidance answers are more similar to their questions w.r.t. sequential form

**2.2 ‘Fight’ vs. ‘Flight’**

* Average Levenshtein distance for 'Fight' Q-A pairs:
  + Threshold 50: 302
    - 10% percentile: 153.2
    - 90% percentile: 417.4
  + No threshold: 403.53
    - 10% percentile: 124.3
    - 90% percentile: 637.2000000000002
* Average Levenshtein distance for 'Flight' Q-A pairs:
  + Threshold 20: 212.29
    - 10% percentile: 38.400000000000006
    - 90% percentile: 428.6
  + No threshold: 375.96
    - 10% percentile: 136.1
    - 90% percentile: 599.2

→ ‘Fight’ pairs are less similar w.r.t. their surface form