However, how do we define p(w|REF)?

- p(w|REF): Reference Language Model
- What do we know about those unseen words?
- Why are there unseen words?
  - Zipf's law: most words occur infrequently in text (e.g., just once)
  - Unseen words are non-relevant to a topic
  - Unseen words are relevant, but the text data sample isn't large enough to include them
- The context variable C in $p(w_1 w_2 \ldots w_m | C)$ can provide a basis for defining p(w|REF)
  - E.g., in retrieval, p(w|Collection) can serve as p(w|REF) for estimating a language model for an individual document p(w|d)