17/4

N-gram	count
your	883,614
rights	80,891
doorposts	21
your rights	378
your doorposts	0

corpus $520 \cdot 10^6$ words $= N$

a) stima con MLE di $P(\text{your}) = \frac{883{,}614}{520 \cdot 10^6}$

$Pr(\text{right} \mid \text{your}) = \frac{378}{883{,}614} \left(= \frac{C(w_i w_{i+1})}{C(w_i *)} \right)$

b) stima $Pr(\text{doorposts} \mid \text{your})$; MLE con add-$k$ smoothing per $k = 0.01$; $|V| = 1{,}254{,}193$

$Pr = \frac{0 + 0.01}{883{,}614 + |V| \cdot k}$ $\Rightarrow$ $C(\text{doorposts your})$ aumenta di k

$\Downarrow$

$C(w)$ aumenta di k $\forall w \in$ corpus

Argomento extra (non in syllabus)
SINTASSI: regole che governano struttura di frasi -> determinata da mente
CONSTITUENT/PHRASE: gruppo di parole che fungono da unità in struttura gerarchica
(es. "He saw the house on the hill" -> "He", "the house", "on the hill")
Per identificarli, CONSTITUENCY TESTS -> più importanti: noun phrase, verb phrase
Parte fondamentale di NLP: AMBIGUITY -> ambiguità in syntactic parsing che influenza interpretazione: problema di PP ATTACHMENT
Si può rappresentare struttura sintattica come albero