Choosing the right *n*

- Choosing *n* is a trade off between:
  - Modeling capacity (increases with *n*)

Examples from Eisenstein (2018)

Gorillas always like to groom **their** friends.

The **computer** that's on the 3rd floor of our office building **crashed**.

  - Tractable estimation (harder with larger *n* … sparsity)

Example from Jurafsky & Martin

	i	want	to	eat	chinese	food	lunch	spend
i	5	827	0	9	0	0	0	2
want	2	0	608	1	6	6	5	1
to	2	0	4	686	2	0	6	211
eat	0	0	2	0	16	2	42	0
chinese	1	0	0	0	0	82	1	0
food	15	0	15	0	1	4	0	0
lunch	2	0	0	0	0	1	0	0
spend	1	0	1	0	0	0	0	0

Figure 3.1 Bigram counts for eight of the words (out of $V = 1446$) in the Berkeley Restaurant Project corpus of 9332 sentences. Zero counts are in gray.

Language model evaluation

- When dealing with BOW features, our input dimensions were fixed $(M \times N)$, giving us a direct means to evaluate our model on any given example i using the NLL:

$$NLL(i^{th}\text{ example}) = -\log P(y^{(i)} \mid \mathbf{x}^{(i)})$$

- In language modeling, $y^{(i)}$ represents the last word in a length T sequence, $\mathbf{x}^{(T)_i}$, while $\mathbf{x}^{(i)}$ is replaced by the prefix of that word: $\{\mathbf{x}^{(1)_i} \ldots \mathbf{x}^{(T-1)_i}\}$. In this case our interpretation of log likelihood is muddled by the fact that, all else being equal, longer sequences yield lower NLL:

$$NLL(i^{th}\text{ example}) = -\sum_{t=1}^{T} \log P(\mathbf{x}^{(t)_i} \mid \mathbf{x}^{(t-n)_i}, \ldots, \mathbf{x}^{(t-1)_i})$$

- *Perplexity offers a useful* evaluation metric for LMs (lower is better):

$$PPL(\mathbf{X}^{(i)}) = \exp\left\{-\frac{1}{T}\sum_{t=1}^{T} \log P\left(\mathbf{x}^{(t)_i} \mid \mathbf{x}^{(t-n)_i} \ldots \mathbf{x}^{(t-1)_i}\right)\right\}$$