Which model is better?

The best *n* might depend on how much data you have:

- bigrams might be not enough
- 7-grams might never occur

Extrinsic evaluation:

- Quality of a downstream task: machine translation, speech recognition, spelling correction…

Intrinsic evaluation:

- Hold-out (text) perplexity!