

# Modeling Grammaticality

[mostly a blackboard lecture]

## Word trigrams: Which sentences are grammatical? A good model of English?

- ✓ [Gen 4:25] And Adam gave names to all feudal, patriarchal, idyllic relations. It has but established new classes, new conditions of oppression, new forms of struggle in place of the West?
- ✓ The bourgeoisie keeps more and more splitting up into two great lights; the greater light to rule the day of my house is this Eliezer of Damascus.
- ? [Gen 14:1] And it came to pass through, they always and everywhere represent the interests of the conditions of oppression, new forms of struggle in place of Sichem, unto the plain of Moreh. And the LORD had respect unto Abel and Abel was a keeper of sheep, but Cain was a sweet finish, after the flood:
- ? [Gen 7:81] Of clean beasts, and of beasts that are not clean by two, the male and his house with great plagues because of the ground of the same character, into one nation, with one government, one code of laws, one national class interest, one frontier, and one custom tariff.
- x The bourgeoisie, possesses, however, this distinct feature: it has simplified class antagonisms. It has been republished in that land, and I will make him drink wine, and we went through it together once more before it went to press. It is entitled: Manifesto of the Communist League, an international

no main verb

## Why it does okay ...

- We never see "the go of" in our *training text*.
- So our dice will never generate "the go of."
  - That trigram has probability 0.

## Why it does okay ... but isn't perfect.

- We never see "the go of" in our *training text*.
- So our dice will never generate "the go of."
  - That trigram has probability 0.
- But we still got some ungrammatical sentences ...
  - All their 3-grams are "attested" in the training text, but still the sentence isn't good.

You shouldn't eat these chickens  
because these chickens eat  
arsenic and bone meal ...

3-gram model

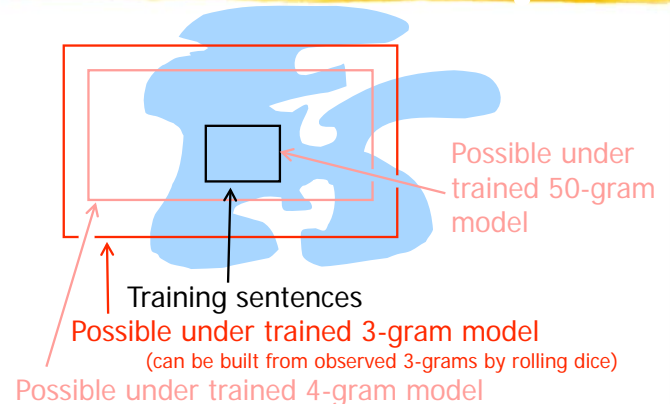
Training sentences

... eat these chickens eat ...

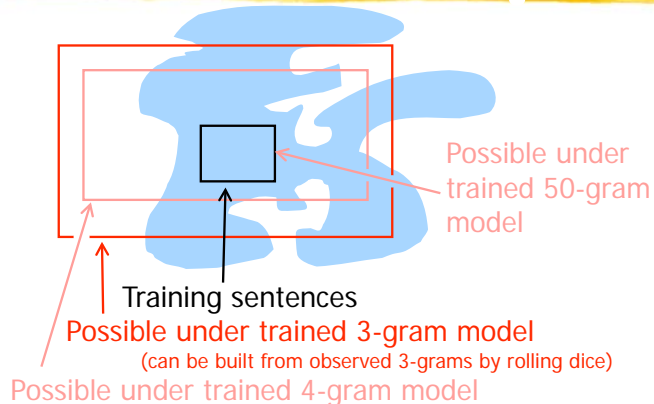
## Why it does okay ... but isn't perfect.

- We never see "the go of" in our *training text*.
- So our dice will never generate "the go of."
  - That trigram has probability 0.
- But we still got some ungrammatical sentences ...
  - All their 3-grams are "attested" in the training text, but still the sentence isn't good.
- Could we rule these bad sentences out?
  - 4-grams, 5-grams, ... 50-grams?
  - Would we now generate only grammatical English?

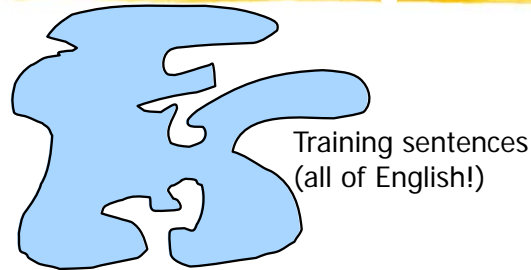
## Grammatical English sentences



What happens as you increase the amount of training text?



What happens as you increase the amount of training text?



Now where are the 3-gram, 4-gram, 50-gram boxes?  
 Is the 50-gram box now perfect?  
 (Can any model of language be perfect?)  
 Can you name some non-blue sentences in the 50-gram box?

Are n-gram models enough?

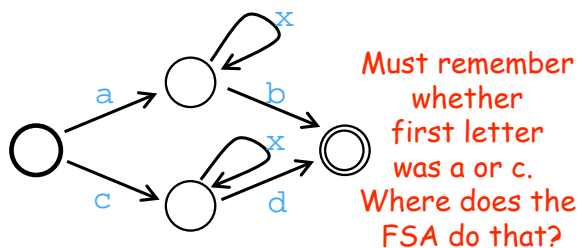
- Can we make a list of (say) 3-grams that combine into all the grammatical sentences of English?
- Ok, how about only the grammatical sentences?
- How about all and only?

Can we avoid the systematic problems with n-gram models?

- Remembering things from arbitrarily far back in the sentence
  - Was the subject singular or plural?
  - Have we had a verb yet?
- Formal language equivalent:
  - A language that allows strings having the forms  $a x^* b$  and  $c x^* d$  ( $x^*$  means "0 or more  $x$ 's")
  - Can we check grammaticality using a 50-gram model?
  - No? Then what can we use instead?

Finite-state models

- Regular expression:  $a x^* b \mid c x^* d$
- Finite-state acceptor:



Context-free grammars

- ~~Sentence~~  $\rightarrow$  ~~Noun Verb Noun~~
- $S \rightarrow N V N$
- $N \rightarrow \text{Mary}$
- $V \rightarrow \text{likes}$
- How many sentences?
- Let's add:  $N \rightarrow \text{John}$
- Let's add:  $V \rightarrow \text{sleeps}$ ,  $S \rightarrow N V$

## Write a grammar of English

- You have 2 weeks. 😊



### Syntactic rules.

- 1  $S1 \rightarrow NP VP .$
- 1  $VP \rightarrow VerbT NP$
- 20  $NP \rightarrow Det N'$
- 1  $NP \rightarrow Proper$
- 20  $N' \rightarrow Noun$
- 1  $N' \rightarrow N' PP$
- 1  $PP \rightarrow Prep NP$

## 3. Now write a grammar of English

### Lexical rules.

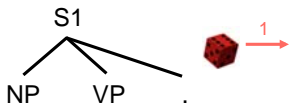
- 1 Noun  $\rightarrow$  castle
- 1 Noun  $\rightarrow$  king
- ...
- 1 Proper  $\rightarrow$  Arthur
- 1 Proper  $\rightarrow$  Guinevere
- ...
- 1 Det  $\rightarrow$  a
- 1 Det  $\rightarrow$  every
- ...
- 1 VerbT  $\rightarrow$  covers
- 1 VerbT  $\rightarrow$  rides
- ...
- 1 Misc  $\rightarrow$  that
- 1 Misc  $\rightarrow$  bloodier
- 1 Misc  $\rightarrow$  does
- ...

### Syntactic rules.

- 1  $S1 \rightarrow NP VP .$
- 1  $VP \rightarrow VerbT NP$
- 20  $NP \rightarrow Det N'$
- 1  $NP \rightarrow Proper$
- 20  $N' \rightarrow Noun$
- 1  $N' \rightarrow N' PP$
- 1  $PP \rightarrow Prep NP$

Any PCFG is okay

## 3. Now write a grammar of English



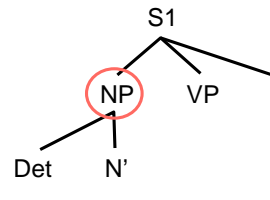
### Here's one to start with.

- 1  $S1 \rightarrow NP VP .$
- 1  $VP \rightarrow VerbT NP$
- 20  $NP \rightarrow Det N'$
- 1  $NP \rightarrow Proper$
- 20  $N' \rightarrow Noun$
- 1  $N' \rightarrow N' PP$
- 1  $PP \rightarrow Prep NP$

Sample a sentence on the blackboard

Any PCFG is okay

## 3. Now write a grammar of English



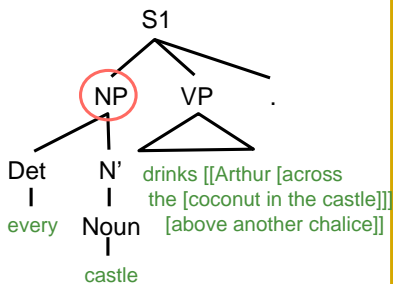
### Here's one to start with.

- 1  $S1 \rightarrow NP VP .$
- 1  $VP \rightarrow VerbT NP$
- 20  $NP \rightarrow Det N'$
- 1  $NP \rightarrow Proper$
- 20  $N' \rightarrow Noun$
- 1  $N' \rightarrow N' PP$
- 1  $PP \rightarrow Prep NP$

Sample a sentence on the blackboard

Any PCFG is okay

## 3. Now write a grammar of English



Sample a sentence on the blackboard

### Here's one to start with.

- 1  $S1 \rightarrow NP VP .$
- 1  $VP \rightarrow VerbT NP$
- 20  $NP \rightarrow Det N'$
- 1  $NP \rightarrow Proper$
- 20  $N' \rightarrow Noun$
- 1  $N' \rightarrow N' PP$
- 1  $PP \rightarrow Prep NP$

Arbitrary PCFG is okay