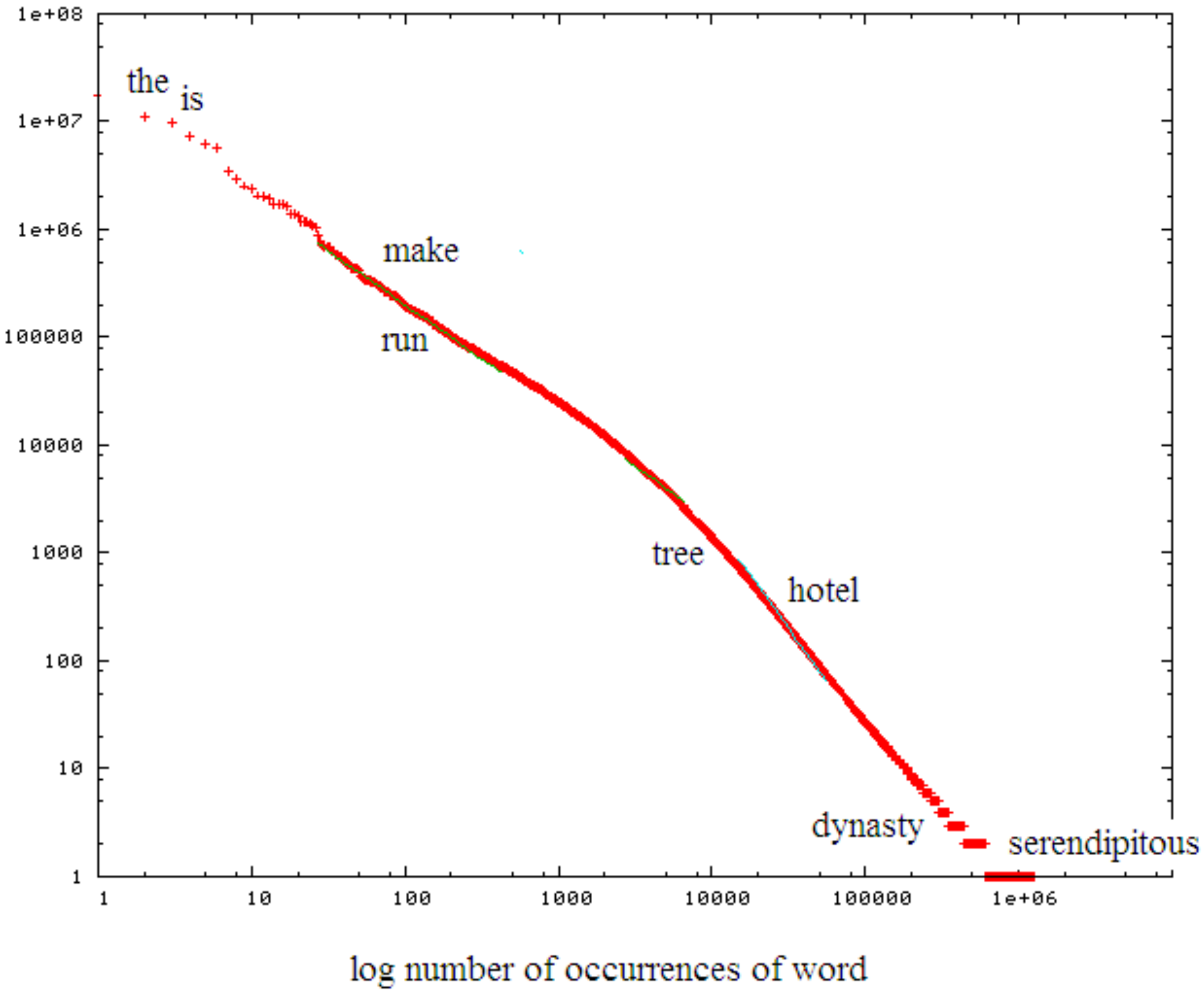
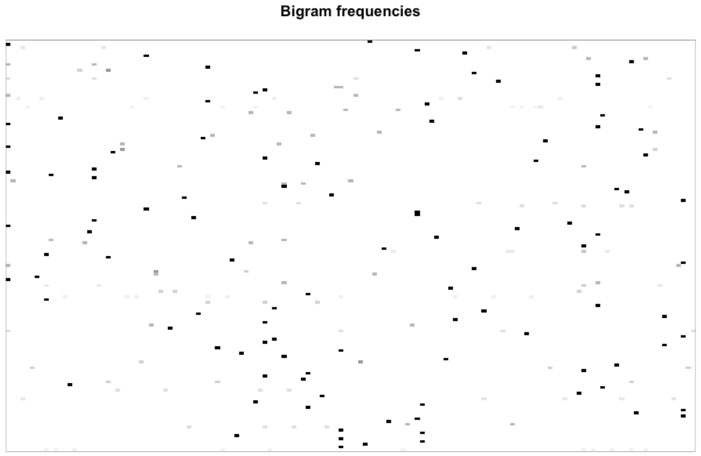
Computational Cognitive Science



Lecture 16 and 17: Sequential learning with n-grams

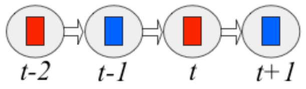


	Red	Blue
Red	$p(R\|R)$	$p(R\|B)$
Blue	$p(B\|R)$	$P(B\|B)$

T

X_t = colour at time t

$S = \{R, B\}$



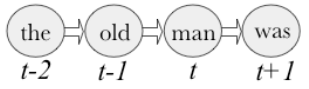
$P(w_2|w_1)$ w_1

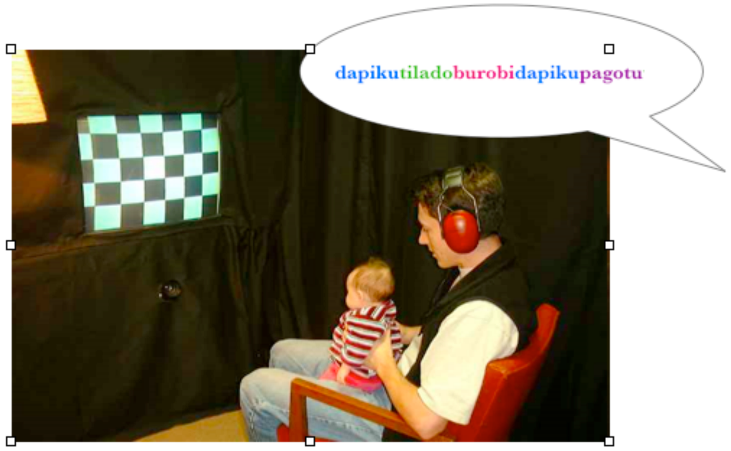
w_2

	The	Man	Ate	Old	Fruit	Who	Was
The	0	0	1.0	0	0	0	1.0
Man	0.33	0	0	1.0	0	0	0
Ate	0	0	0	0	0	1.0	0
Old	0.33	0	0	0	0	0	0
Fruit	0.33	0	0	0	0	0	0
Who	0	0.5	0	0	0	0	0
Was	0	0.5	0	0	0	0	0

X_t = word at time t

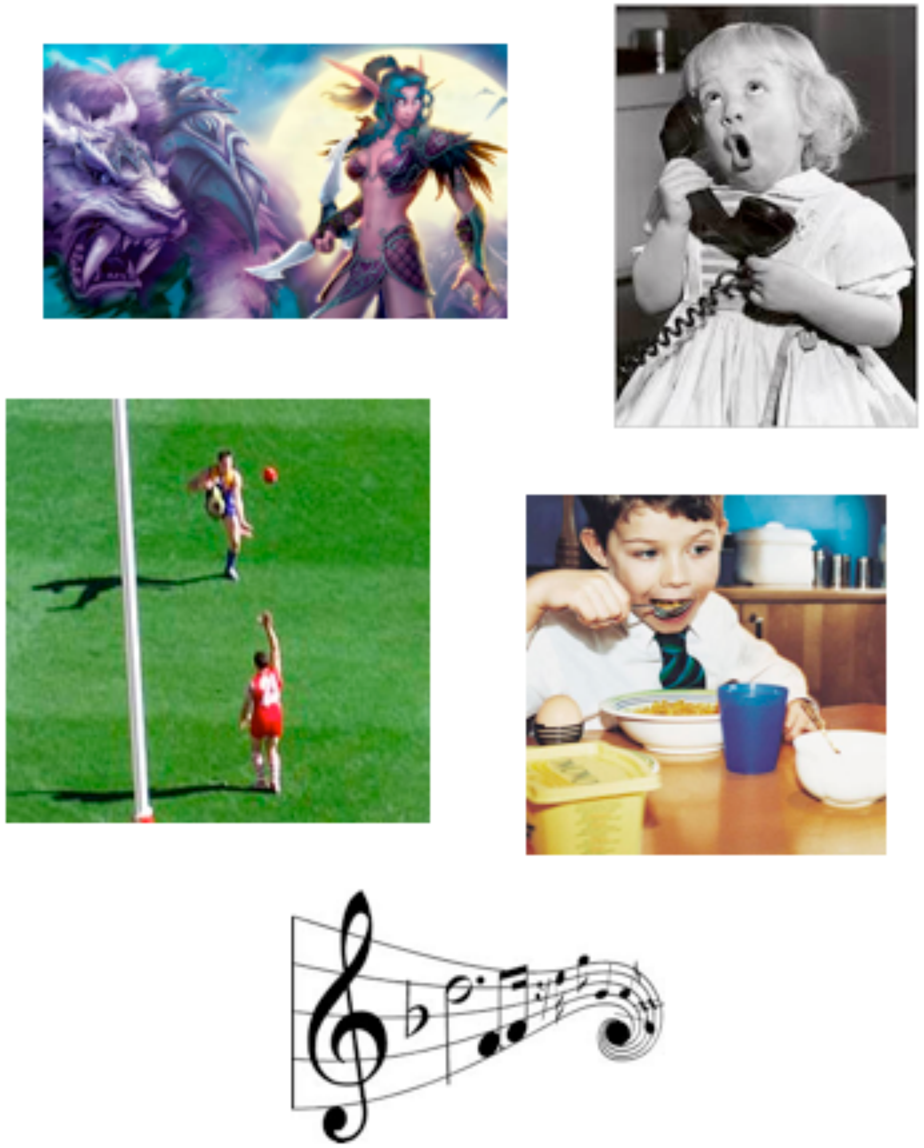
$S = \{the, old, man, was, \ldots\}$





Last time

- Sequence learning is an important problem in cognition, and language is a clear example of when this is relevant

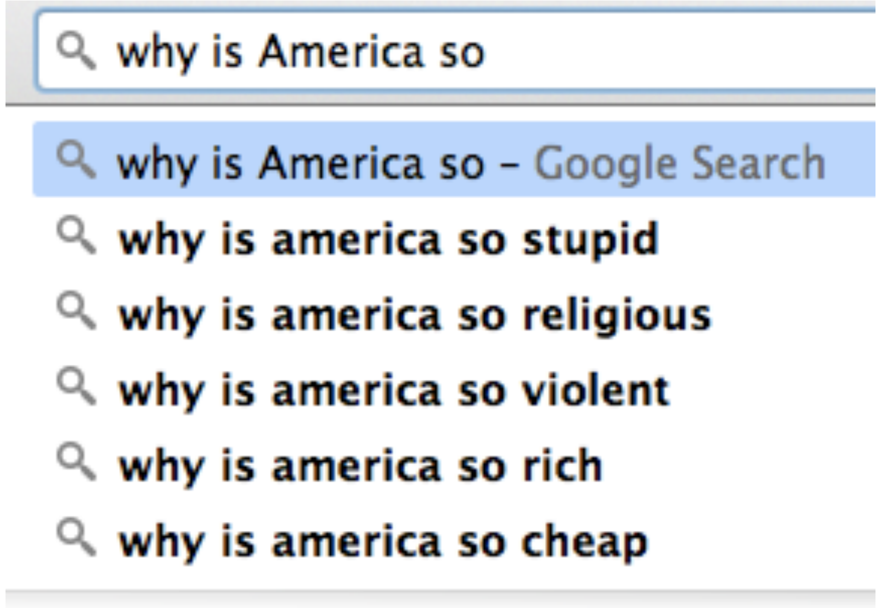


Last time

- Sequence learning is an important problem in cognition, and language is a clear example of when this is relevant
- n-gram models, which calculate the probability of an item given the previous n-1 items, are widely used in natural language processing to address this problem.

why is Australia so

why is Australia so – Google Search
why is australia so expensive
why is australia so hot
why is australia so great
why is australia so dry
why is australia so boring



why is America so

why is America so – Google Search
why is america so stupid
why is america so religious
why is america so violent
why is america so rich
why is america so cheap

Original plan for the lectures

- Monday: a simple model for sequence learning (n-grams)
  - Description of the approach
  - Application to natural language processing
  - The problem of overfitting
- Tuesday 1: applications of n-gram models
  - A solution to the problem of overfitting
  - Word segmentation
  - Nonadjacent learning
  - What about more complex structure?
- Tuesday 2: extending n-grams (HMMs)
  - Computing likelihood of observations
  - Inferring the hidden state sequence
  - Finding the best HMM (if time)

New plan for the lectures

- Yesterday: a simple model for sequence learning (n-grams)
  - Application to natural language processing
- Today: n-gram models
  - Description of the approach
  - The problem of overfitting
  - A solution to the problem of overfitting
  - Some applications
- After mid-semester break: extending n-grams (HMMs)
  - What about more complex structure?
  - Computing likelihood of observations
  - Inferring the hidden state sequence
  - Finding the best HMM (if time)

New plan for the lectures

- Yesterday: a simple model for sequence learning (n-grams)
  - Application to natural language processing
- Today: n-gram models
  - Description of the approach
  - The problem of overfitting
  - A solution to the problem of overfitting
  - Some applications
- After mid-semester break: extending n-grams (HMMs)
  - What about more complex structure?
  - Computing likelihood of observations
  - Inferring the hidden state sequence
  - Finding the best HMM (if time)

N-grams: tracking clusters of words

- For both generation and prediction, higher n is better!
- Both are extremely straightforward given the n-gram probabilities

Two kinds of probabilities

1. Probability of a word or series of words

raw: $p(w_1, ..., w_n)$

2. Probability of a word given a previous word or series of words

conditional: $p(w_n | w_1, ...w_{n-1})$

The equations are distinct (except in the unigram case)

Raw probability of n words: $P(w_1,\ldots,w_n)$

Simplest way to calculate this:
Maximum Likelihood Estimation (MLE)
based on observed frequencies

$$p(w_1, \ldots, w_n) = \frac{C(w_1, \ldots w_n)}{N}$$

Count C of times $w_1 \ldots w_n$ is in the corpus

Probability of observing the series of words $w_1 \ldots w_n$

# of training instances (i.e., total different series of n words possible in that corpus)

Raw probability of n words: $P(w_1,\ldots,w_n)$

If n=1, this reduces to the frequency of each word!

$$p(w_n) = \frac{C(w_n)}{N}$$

Count C of times w_n is in the corpus

Probability of observing the series of words w_n

# of training instances (i.e., total words in the corpus: that is, corpus length, not vocabulary size)

Raw probability of n words: $P(w_1,\ldots,w_n)$

If n=1, this reduces to the frequency of each word!

The old man was the man who ate the fruit.

P(the) = 3/10 = 0.3
P(man) = 2/10 = 0.2
P(ate) = 1/10 = 0.1
P(old) = 1/10 = 0.1
P(who) = 1/10 = 0.1
P(fruit) = 1/10 = 0.1
P(was) = 1/10 = 0.1

Raw probability of *n* words: $P(w_1,...,w_n)$

If $n>1$, it is important to make sure the N in the denominator is the total number of n-grams (of that n) in the corpus

This is generally less useful than being able to predict things..

The old man was the man who ate the fruit.

9 total bigrams in the corpus:

the old
old man
man was
was the
the man
man who
who ate
ate the
the fruit

P(the ate) = 0/9
P(the fruit) = 1/9
P(the man) = 1/9
P(the old) = 1/9
P(the the) = 0/9
P(the was) = 0/9
P(the who) = 0/9
P(ate fruit) = 0/9
…

Conditional: predicting the next word: $P(w_n|w_{1},...,w_{n-1})$

The MLE probability of a word given a previous word or series of words is given by:

$$p(w_n|w_1, ...w_{n-1}) = \frac{C(w_1,...,w_n)}{C(w_1,...w_{n-1})}$$

Count C of times there are n words in a row

Probability of w_n given previous n-1 words

Count C of times of previous n-1 words in a row are observed

Conditional: predicting the next word: $P(w_n|w_1,\ldots,w_{n-1})$

If $n=2$, this means calculating the frequency of each word given one previous other

The old
man was
the man
who ate
the fruit.

$P(w_2|w_1)$

w_1

w_2

	The	Man	Ate	Old	Fruit	Who	Was
The	0	0	1.0	0	0	0	1.0
Man	0.33	0	0	1.0	0	0	0
Ate	0	0	0	0	0	1.0	0
Old	0.33	0	0	0	0	0	0
Fruit	0.33	0	0	0	0	0	0
Who	0	0.5	0	0	0	0	0
Was	0	0.5	0	0	0	0	0

Conditional: predicting the next word: $P(w_n|w_1,\ldots,w_{n-1})$

This is just a Markov chain!

$P(w_2|w_1)$

w_1

w_2

	The	**Man**	**Ate**	**Old**	**Fruit**	**Who**	**Was**
The	0	0	1.0	0	0	0	1.0
Man	0.33	0	0	1.0	0	0	0
Ate	0	0	0	0	0	1.0	0
Old	0.33	0	0	0	0	0	0
Fruit	0.33	0	0	0	0	0	0
Who	0	0.5	0	0	0	0	0
Was	0	0.5	0	0	0	0	0

Conditional: predicting the next word: $P(w_n|w_1,...,w_{n-1})$

This is just a Markov chain!

It is a matrix defining the transition probabilities between a set of states.

	Red	Blue
Red	$p(R\|R)$	$p(R\|B)$
Blue	$p(B\|R)$	$P(B\|B)$

T

$P(w_2|w_1)$

w_1

w_2

	The	Man	Ate	Old	Fruit	Who	Was
The	0	0	1.0	0	0	0	1.0
Man	0.33	0	0	1.0	0	0	0
Ate	0	0	0	0	0	1.0	0
Old	0.33	0	0	0	0	0	0
Fruit	0.33	0	0	0	0	0	0
Who	0	0.5	0	0	0	0	0
Was	0	0.5	0	0	0	0	0

	h_1	h_2	h_3
h_1	$p(h_1\|h_1)$	$p(h_1\|h_2)$	$p(h_1\|h_3)$
h_2	$p(h_2\|h_1)$	$p(h_2\|h_2)$	$p(h_2\|h_3)$
h_3	$p(h_3\|h_1)$	$p(h_3\|h_2)$	$p(h_3\|h_3)$

T

Formal definition of a Markov chain

Let $X = (X_1, \ldots, X_T)$ be a sequence of random variables taking values in the state space: some countable set $S=\{s_1, \ldots, s_N\}$.

	Red	Blue
Red	$p(R\mid R)$	$p(R\mid B)$
Blue	$p(B\mid R)$	$P(B\mid B)$

T

$P(w_2|w_1)$ w_1

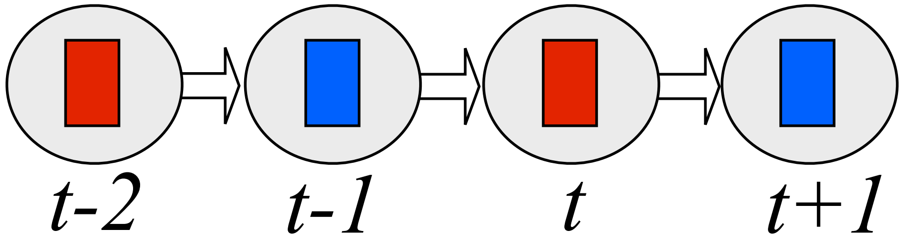
w_2	The	Man	Ate	Old	Fruit	Who	Was
The	0	0	1.0	0	0	0	1.0
Man	0.33	0	0	1.0	0	0	0
Ate	0	0	0	0	0	1.0	0
Old	0.33	0	0	0	0	0	0
Fruit	0.33	0	0	0	0	0	0
Who	0	0.5	0	0	0	0	0
Was	0	0.5	0	0	0	0	0

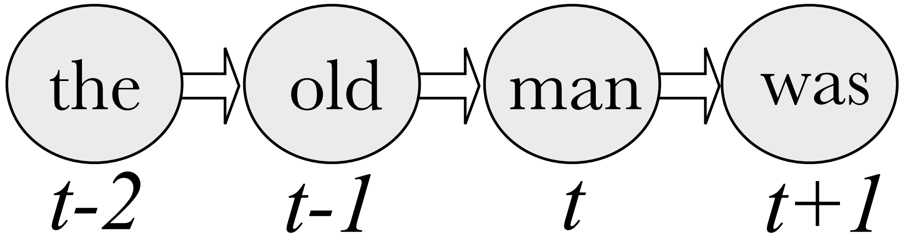
X_t = colour at time t

$S = \{R, B\}$

X_t = word at time t

$S = \{the, old, man, was, \ldots\}$





Formal definition of a Markov chain

Let $X = (X_1, \ldots, X_T)$ be a sequence of random variables taking values in the state space: some countable set $S=\{s_1, \ldots, s_N\}$.

- All Markov chains have the **Markov property**, also called **limited horizon**: the probability of moving into a new state depends only on the current one, not on any previous ones

$$p(X_{t+1} = s_k | X_1, \ldots, X_t) = p(X_{t+1} = s_k | X_t)$$

(For this reason, Markov models are often called **memoryless learners**)

	Red	Blue
Red	0.25	0.75
Blue	0.75	0.25

🟥🟦🟥🟥🟦🟥🟦🟦🟥🟦🟥🟦🟦🟥🟥

The probability of being R at time $t+1$ given the previous colour is the same as the probability of R at $t+1$ given the previous *several* colours

Equivalency to an n-gram?

Does limited horizon mean that Markov Models are equivalent to bigram models only, or are they more generically equivalent to any type of n-gram model?

Answer: they are (or can be) equivalent to any n-gram model, not just bigram models.

Consider what a tri-gram model would look like

If n=3, this means calculating the frequency of each word given two previous others

The old man was the man who ate the fruit.

$P(w_3|w_1,w_2)$

w_1, w_2

w_3

	the ate	the fruit	the old	the man	the the
ate	0	0	0	0	0
fruit	0	0	0	0	0
man	0	0	0	0	0
old	0	0	0	0	0
the	0	0	0	0	0
who	0	0	0	1.0	0
was	0	0	0	0	0

…

…

Consider what a tri-gram model would look like

This is still a matrix of transition probabilities -- the states are just different ones!

Each state now is one of the possible combinations of two words

$P(w_3|w_1,w_2)$

w_1, w_2

	the ate	the fruit	the old	the man	the the
ate	0	0	0	0	0
fruit	0	0	0	0	0
man	0	0	0	0	0
old	0	0	0	0	0
the	0	0	0	0	0
who	0	0	0	1.0	0
was	0	0	0	0	0

	Red	Blue
Red	$p(R\|R)$	$p(R\|B)$
Blue	$p(B\|R)$	$P(B\|B)$

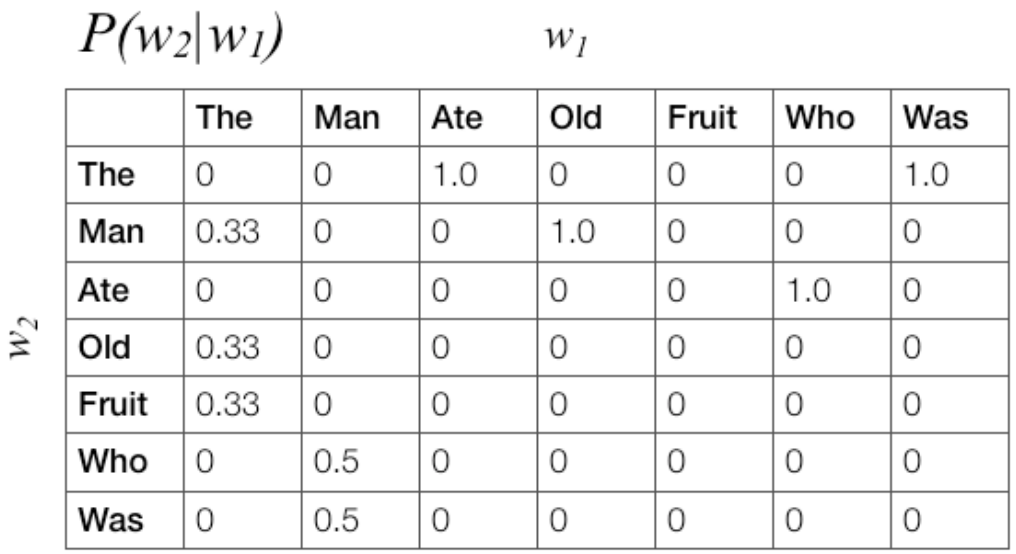
T

Equivalency to n-grams?

Any higher-order n-gram with redescribed states is a Markov model; limited horizon requires only that the states depend on *some* finite number of previous states, not necessarily one

You may have noticed one other thing...

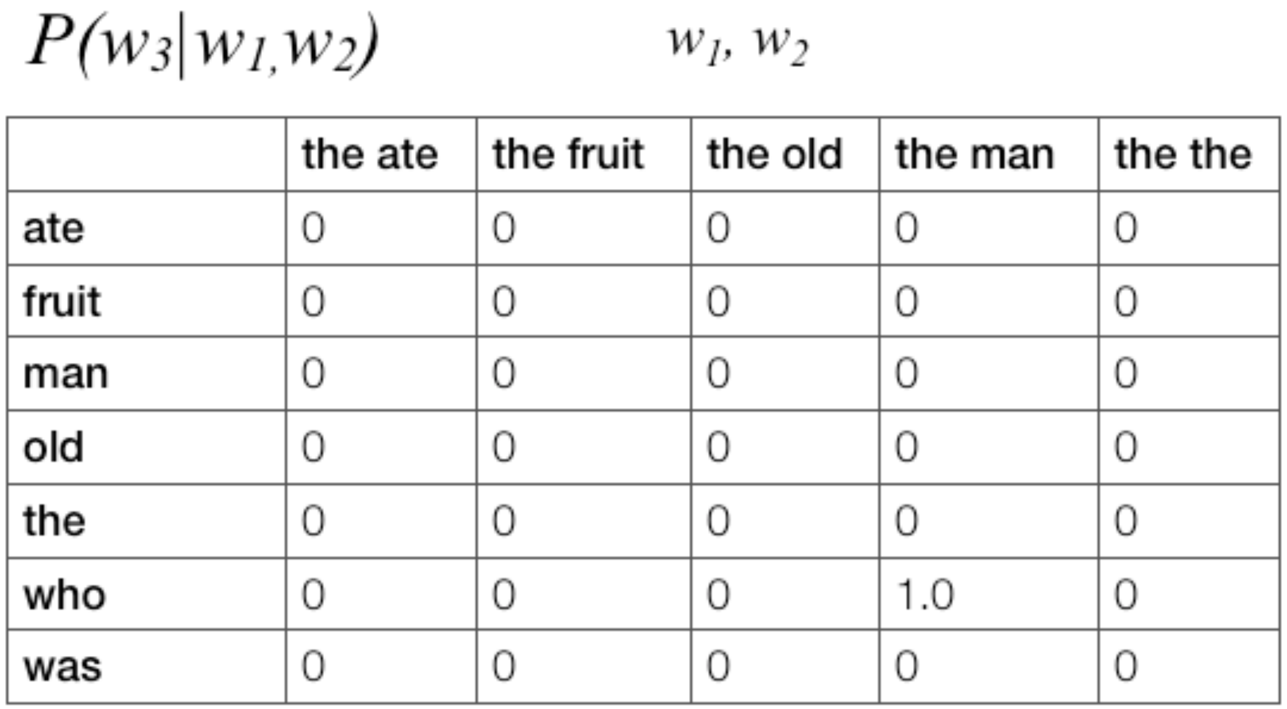
Increasing n causes a huge explosion in the number of states / parameters of the model



$P(w_2|w_1)$ w_1

w_2

	The	Man	Ate	Old	Fruit	Who	Was
The	0	0	1.0	0	0	0	1.0
Man	0.33	0	0	1.0	0	0	0
Ate	0	0	0	0	0	1.0	0
Old	0.33	0	0	0	0	0	0
Fruit	0.33	0	0	0	0	0	0
Who	0	0.5	0	0	0	0	0
Was	0	0.5	0	0	0	0	0



$P(w_3|w_1,w_2)$ w_1, w_2

	the ate	the fruit	the old	the man	the the
ate	0	0	0	0	0
fruit	0	0	0	0	0
man	0	0	0	0	0
old	0	0	0	0	0
the	0	0	0	0	0
who	0	0	0	1.0	0
was	0	0	0	0	0

...

For a vocabulary size V, the number of states is V^n

If V=20,000

unigram: 20000

bigram: $(20000)^2$ = 400 million

trigram: $(20000)^3$ = 8 trillion

4-gram: $(20000)^4$ = 1.6×10^{17}

You may have noticed one other thing...

Increasing n causes a huge explosion in the number of states / parameters of the model

We therefore would like to implement the smallest n-gram that does the job.

For a vocabulary size V, the number of states is V^n

If V=20,000

unigram: 20000
bigram: $(20000)^2$ = 400 million
trigram: $(20000)^3$ = 8 trillion
4-gram: $(20000)^4 = 1.6 \times 10^{17}$

Implementing an n-gram model

No big trick -- basically just go through and tally the counts

```
Process the code (remove commas, add start/end symbols
Create blank array of bigrams of size nwords x nwords
Create a wordlist of all words, each with an index $i$

For each word $w$ = 2 to end of corpus
    Find the index $i_w$ of that word in the wordlist
    Find the index $i_{w-1}$ of the previous word in the wordlist
    Add 1 count to bigram array at $(i_{w-1}, i_w)$
End

Raw probabilities:
  Normalise bigram array by sum of total counts
Conditional probabilities:
  Normalise bigram array by counts of each individual word
```

Results

First, let's try it on a simple newspaper article

Residents across much of southern Australia are bracing for another heatwave, with temperatures forecast to reach into the 40s in some areas today. $ Total fire bans have been issued across South Australia, Victoria and Tasmania ahead of the extreme heat. $ Adelaide's maximum temperature today is expected to be 41 degrees Celsius, with 40C on Friday, 41C on Saturday and 40C on Sunday. $ A catastrophic fire danger rating has been issued for the state's lower southeast. $ Country Fire Service state coordinator Brenton Eden says the weather conditions in South Australia could not be worse. $ We are
s burning in the state he said. $ Firefighters have
Ranges for a fortnight. $ Victoria is also on fire
lbourne and up to 42C in the state's west. $ The
ng for the South West and Wimmera regions, and
extreme conditions. $ Along with scorching heat,
western Victoria. $ Those conditions would lead to
pokesman Steven Walls said. $ Most of Victoria
tions, so we are asking all Victorians to take
hat might cause fires, that includes machinery. $

Scorching heat to return to southern Australia, total fire bans in place

Updated Tue 28 Jan 2014, 11:48am AEDT

Residents across much of southern Australia are bracing for another heatwave, with temperatures forecast to reach into the 40s in some areas today.

Total fire bans have been issued across South Australia, Victoria and Tasmania ahead of the extreme heat.

Adelaide's maximum temperature today is expected to be 41 degrees Celsius, with 40C on Friday, 41C on Saturday and 40C on Sunday.

A catastrophic fire danger rating has been issued for the state's lower south-east.

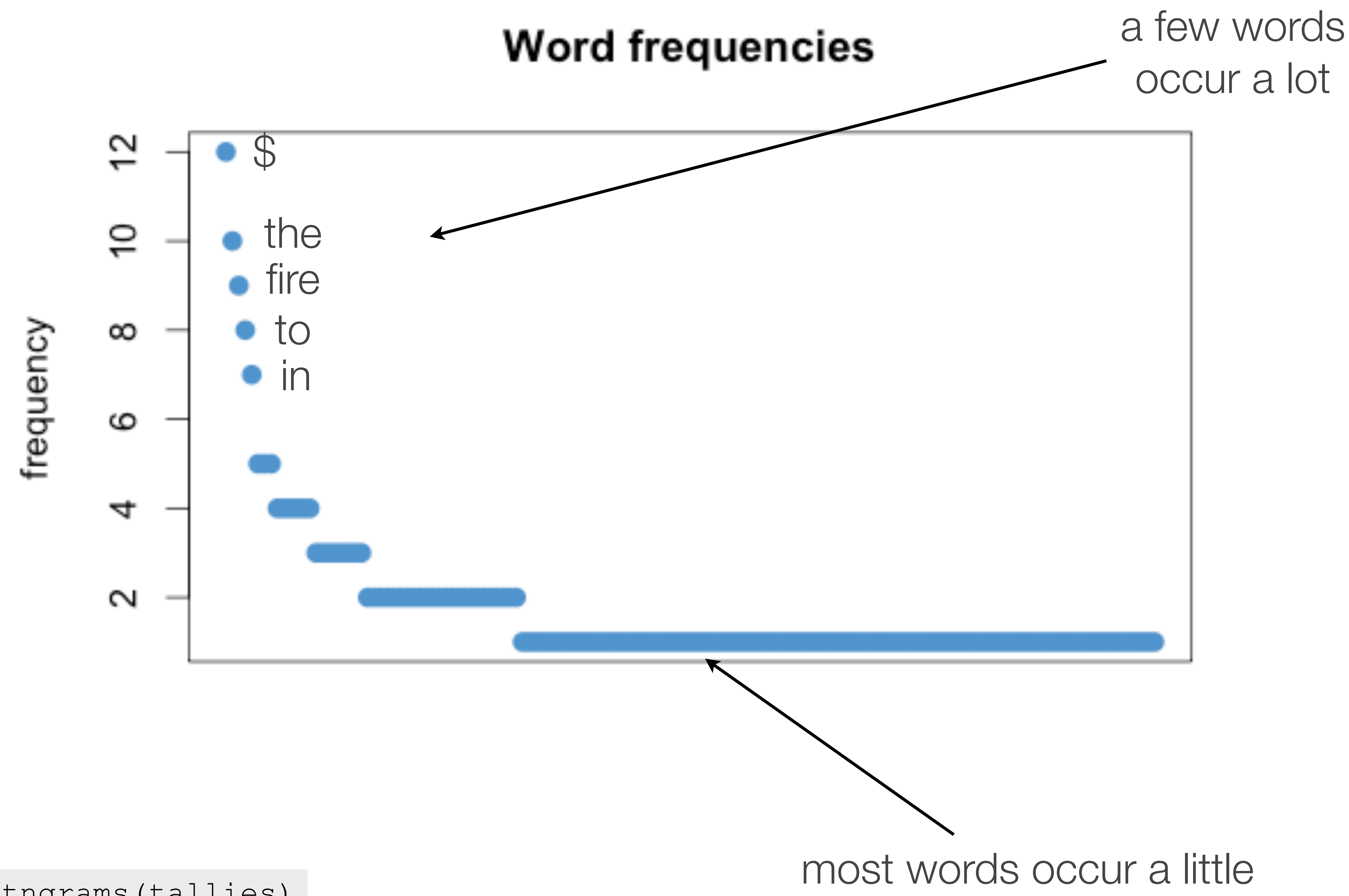
PHOTO: Tmperatures forecast to reach into the 40s in some areas today. (ABC News: Amy Simmons)

Insert these to indicate ends of sentences

```
tallies <- getbigramtallies("weather.txt")
```

http://www.abc.net.au/news/2014-01-28/southern-australia-to-swelter-as-heatwave-returns/5221632

Results

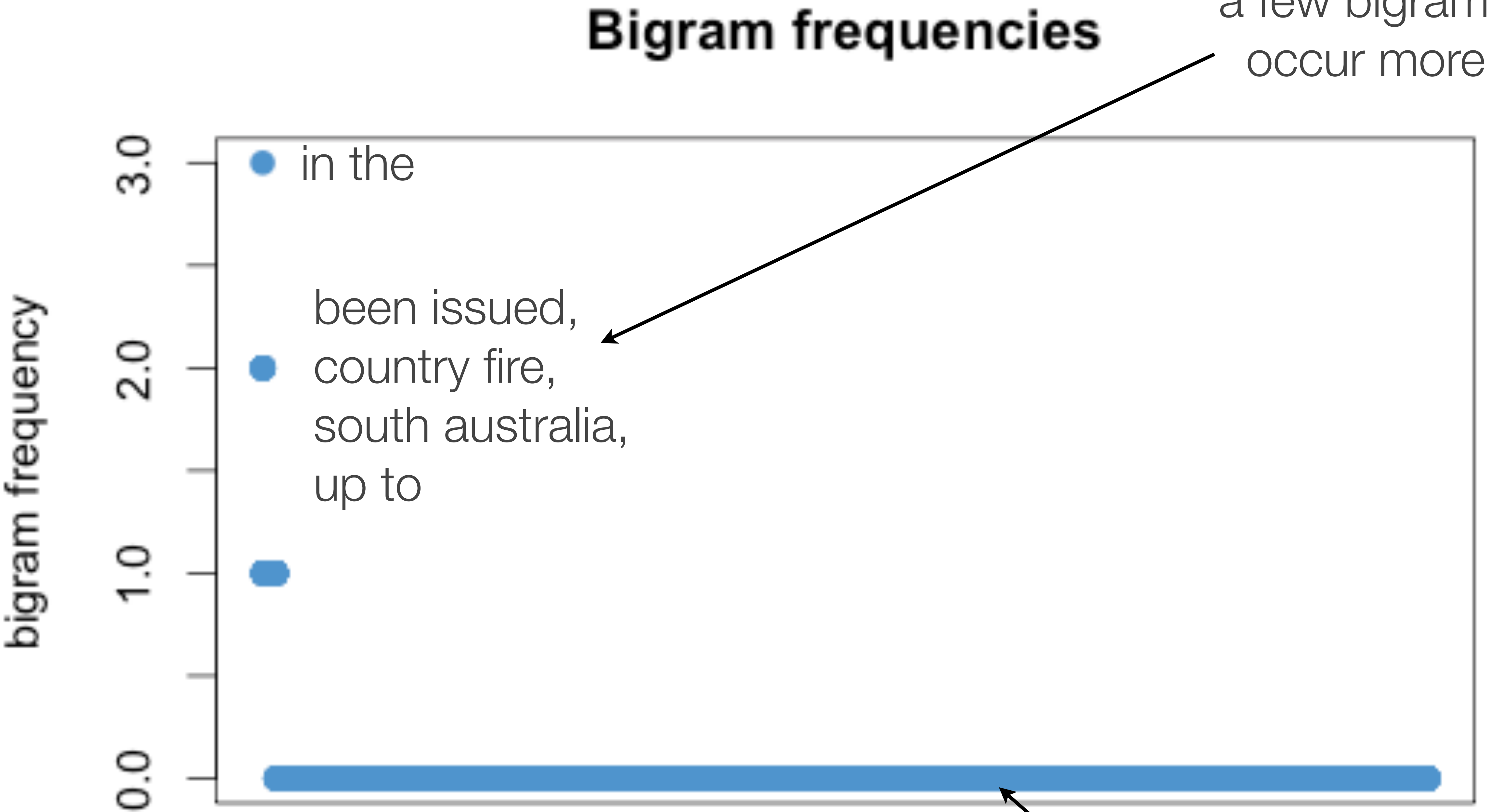


a few words occur a lot

most words occur a little

`plotngrams(tallies)`

Results



in the

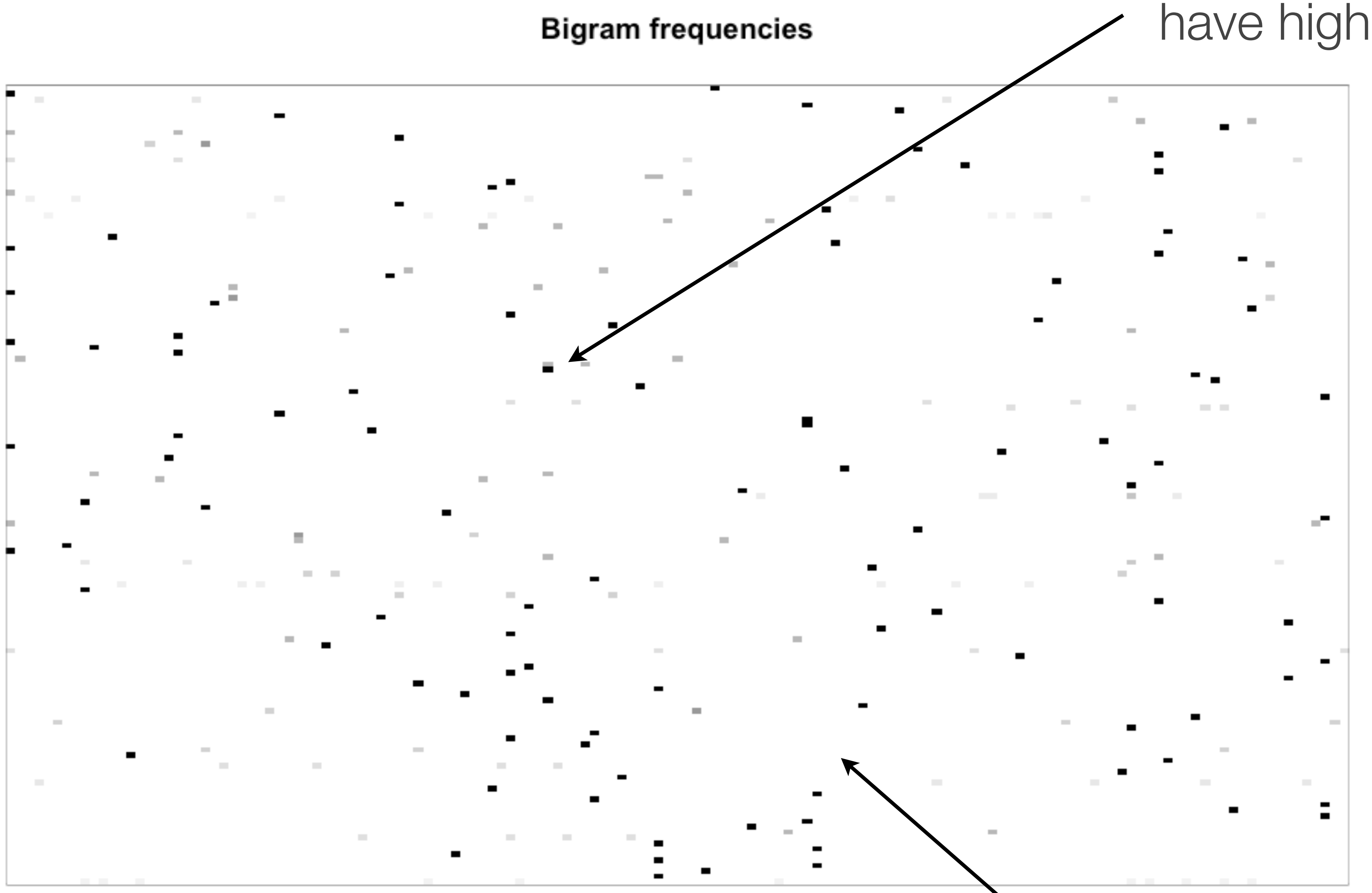
been issued,
country fire,
south australia,
up to

a few bigrams
occur more

most bigrams do not occur

`plotngrams(tallies)`

Results



a few conditional bigrams
have high probability

most bigrams have probability 0

`plotngrams(tallies)`

Results

Most words are low frequency

The bigram matrix is extremely sparse

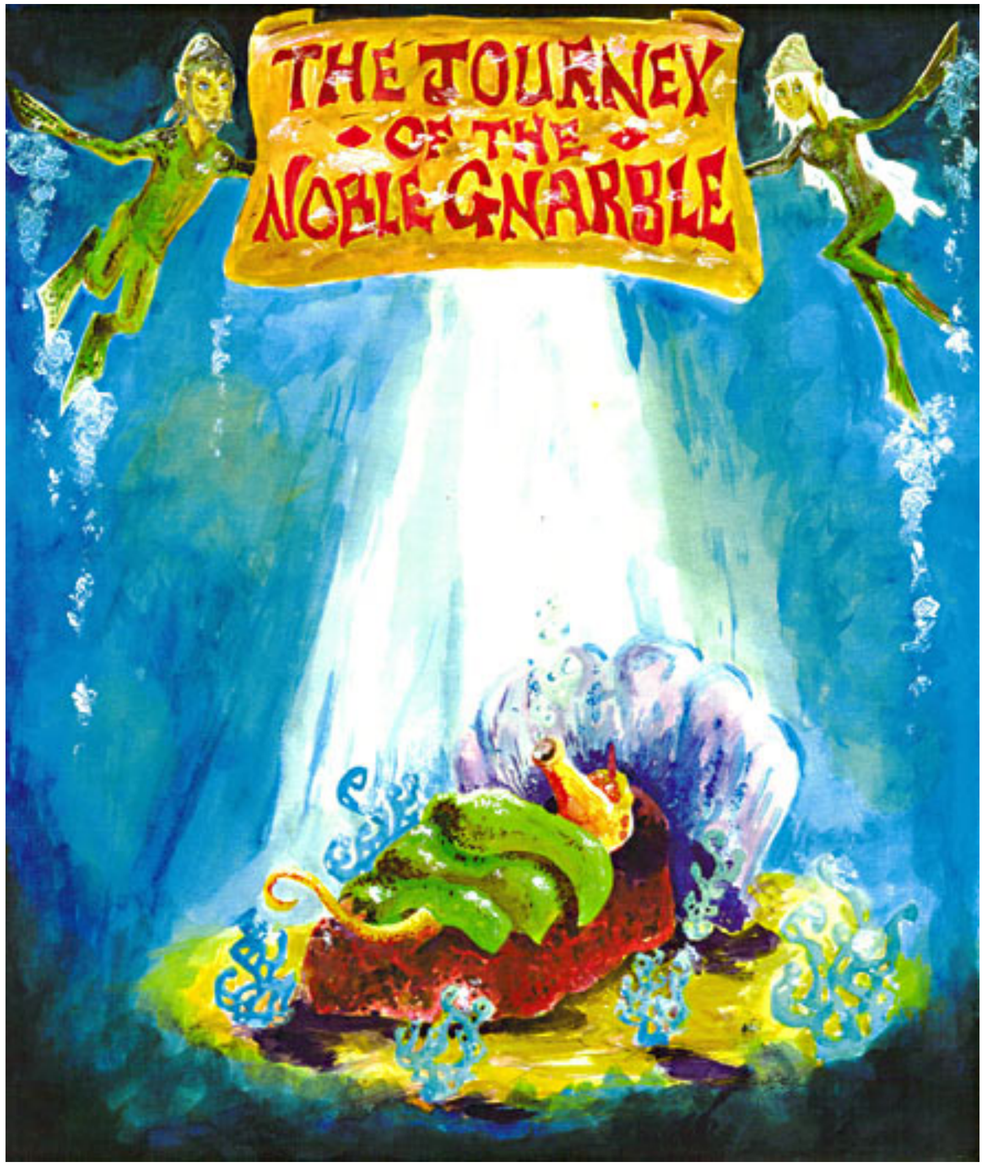
↓

These are related and unavoidable features of language -- not a property of our particular corpus

Results: a different corpus

Online children's book

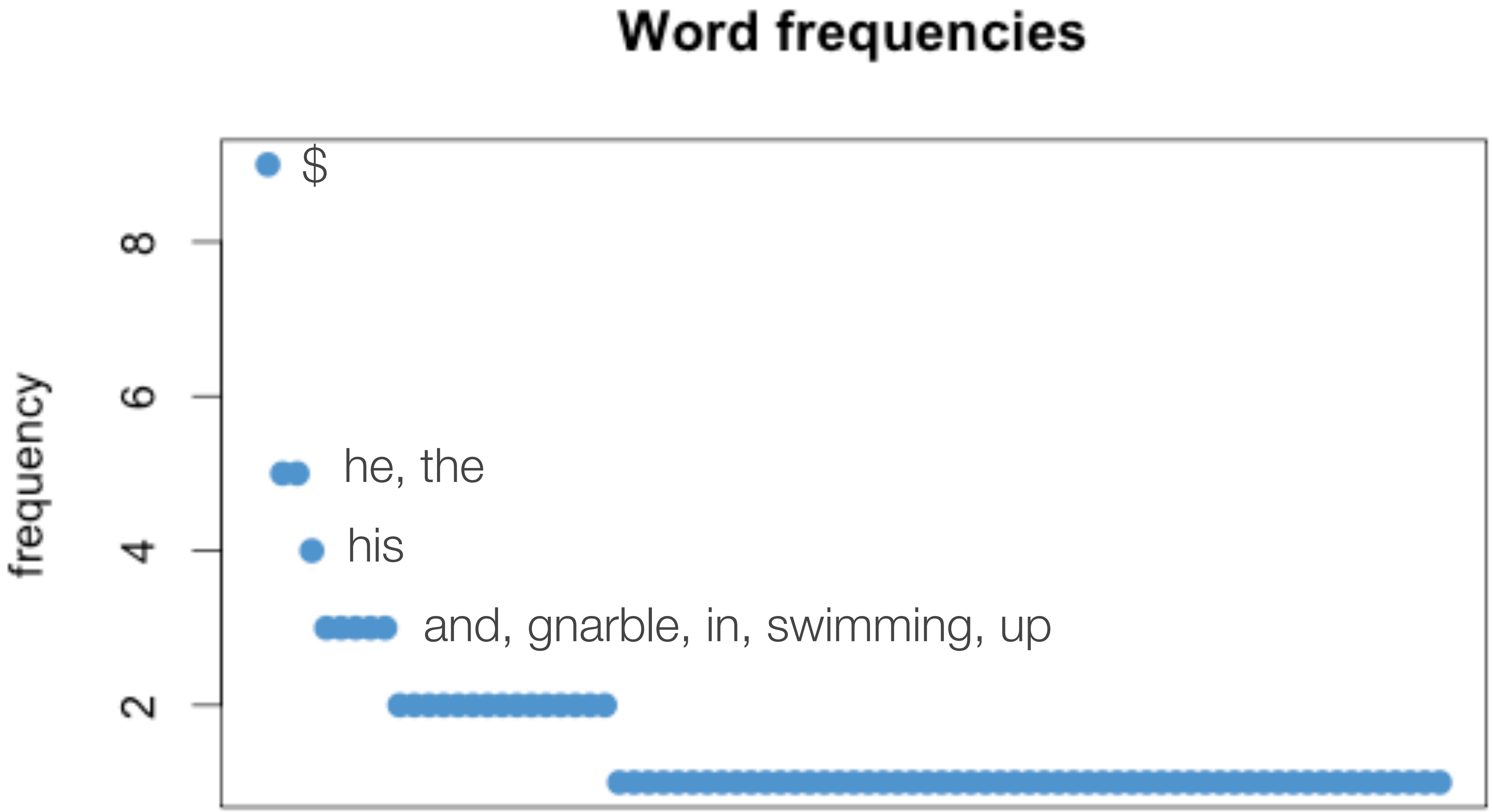
Far below the ocean waves, a gnarble lay in bed. $ All night long his gnarble dreams kept swimming in his head. $ He dreamt a dream of swimming up to see the sky above, lit up by the sun in colors he just knew he'd love. $ But gnarbles never swam that high, their fins were much too small. $ Their tails were thin and floppy, which didn't help at all. $ This gnarble liked his fins and had no problem with his tail. $ So when he woke he knew that he just couldn't, wouldn't fail. $ I'm swimming up above the waves to see the sky of blue. $ I've never seen it even once, and now it's time I do. $



```
tallies <- getbigramtallies("childrensstory.txt")
```

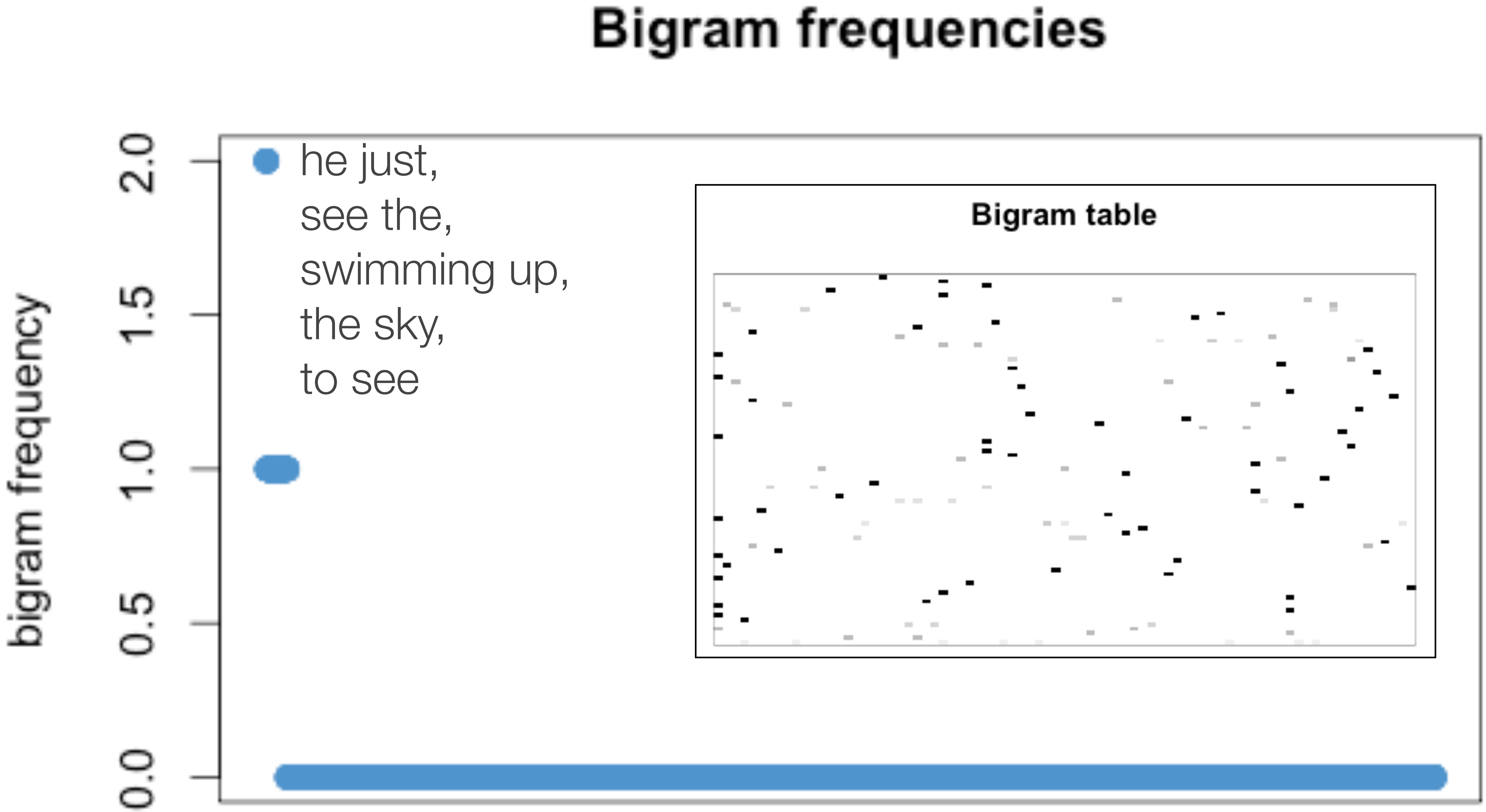
http://www.magickeys.com/books/noblegnarble/

Results: a different corpus



```
plotngrams(tallies)
```

Results: a different corpus



```
plotngrams(tallies)
```

Results: a different corpus

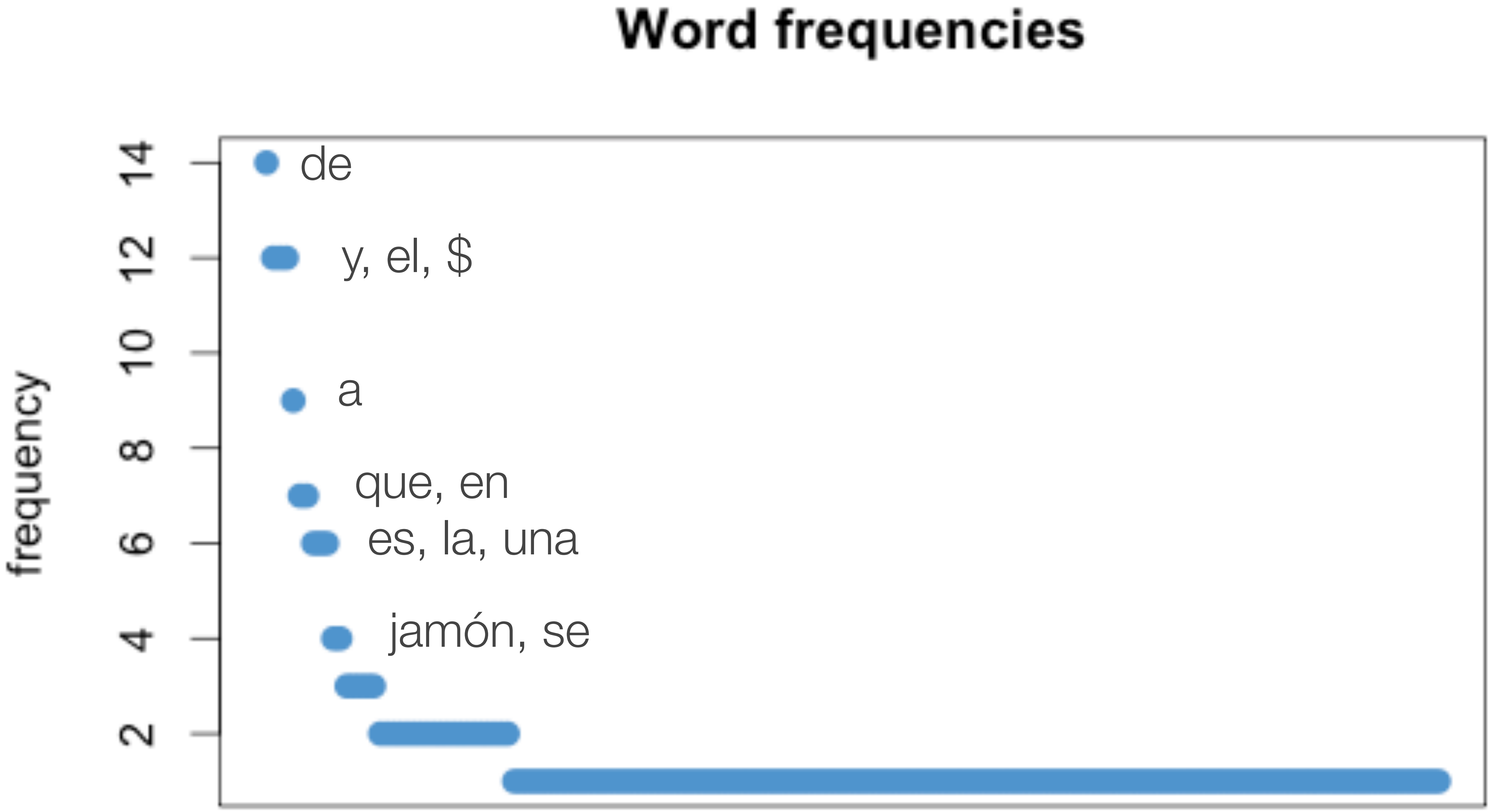
Another language (from an online history of Spanish)

Si algo caracteriza a la gente de Andalucía es su carácter; suelen ser carismáticos, cálidos,
parlanchines y muy amables. $ No reparan en acercarse si los forasteros tienen alguna duda
y, por eso, no fue difícil entablar una conversación con algunos habitantes, quienes al
preguntarles sobre el mejor jamón respondían sin reparo: Hombre, el Cinco Jot
de boca en boca encontramos a María Castro quien es una experta en el tema y
a conocer el génesis del jamón bellotero. $ María se ofreció a pasearnos y llegó
hotel a eso de las once de la mañana para echarnos una mano: Mis amigos dic
como una oficina de turismo de Aracena y es que ese lugar me tiene enamorada
a llevar a conocer todo el proceso, pues ya veréis la diferencia que hay al degu
jamón y cómo uno saborea 103 años de tradición en un solo bocado. $ Sin duda
tenido la fortuna de encontrar a la persona indicada. $ Nos sentamos en el bar
pedimos jamón y una copa de vino tinto. $ Era relativamente temprano para be
cuerpo lo aceptaba gracias al clima fresco y el jugoso embutido. $ Orgullosa
andaluza, María nos explicó que el cerdo ibérico puro es de raza milenaria qu
encuentra en España. $ No hay cerdos como estos en un solo lugar; son autóctonos. $ Es
una mezcla entre el cerdo y el jabalí, tienen la piel oscura, el lomo plano, orejas pequeñas y
encorvadas y patas estrechas de caña con pezuña negra, de ahí que se les diga pata negra.
$ Pero lo más importante es que se alimentan de bellotas, dijo. $

```
tallies <- getbigramtallies("spanishfile.txt")
```

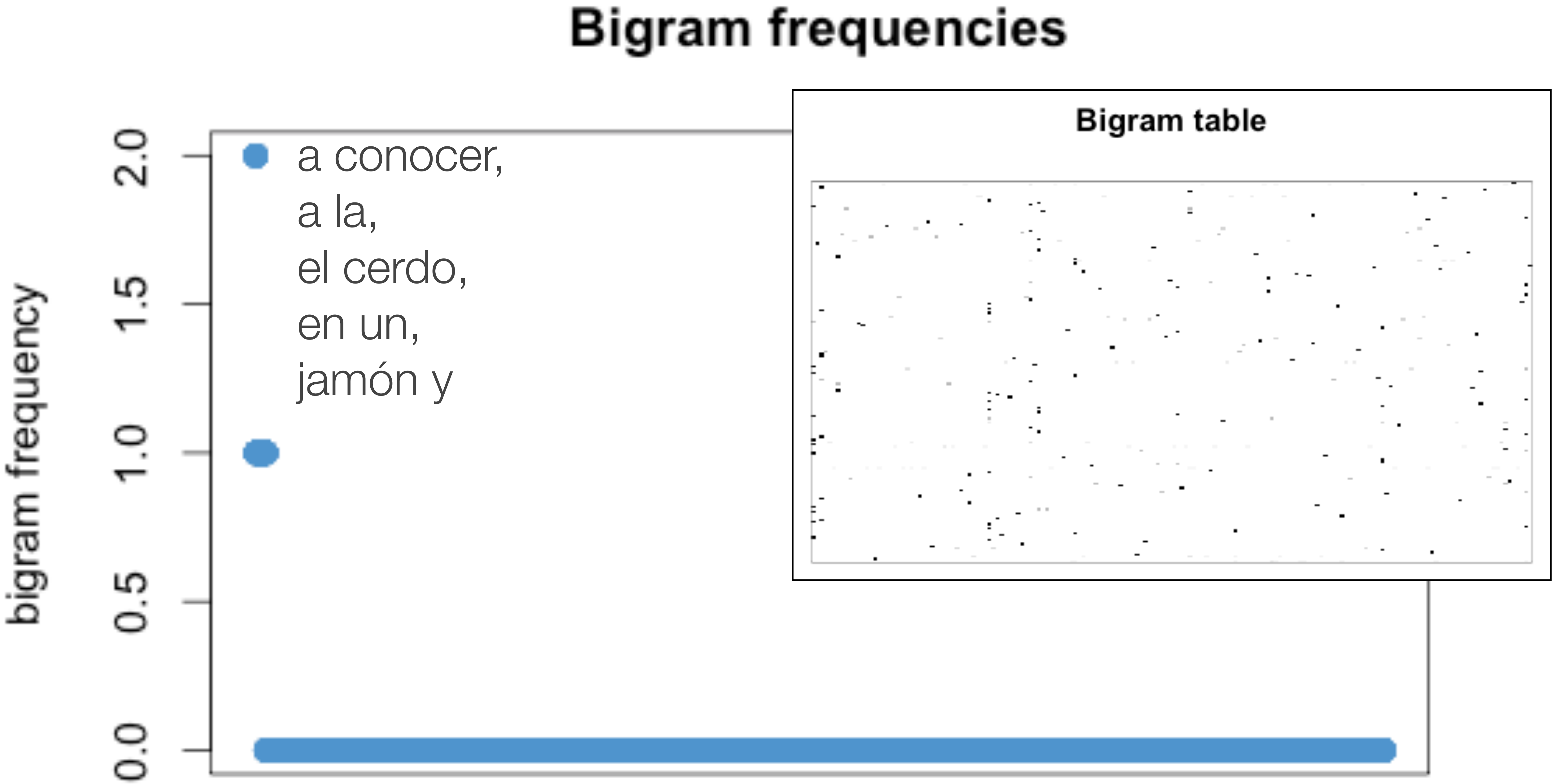
http://www.ngenespanol.com/traveler/gourmet/701340/jamon-jamon/

Results: a different corpus



`plotngrams(tallies)`

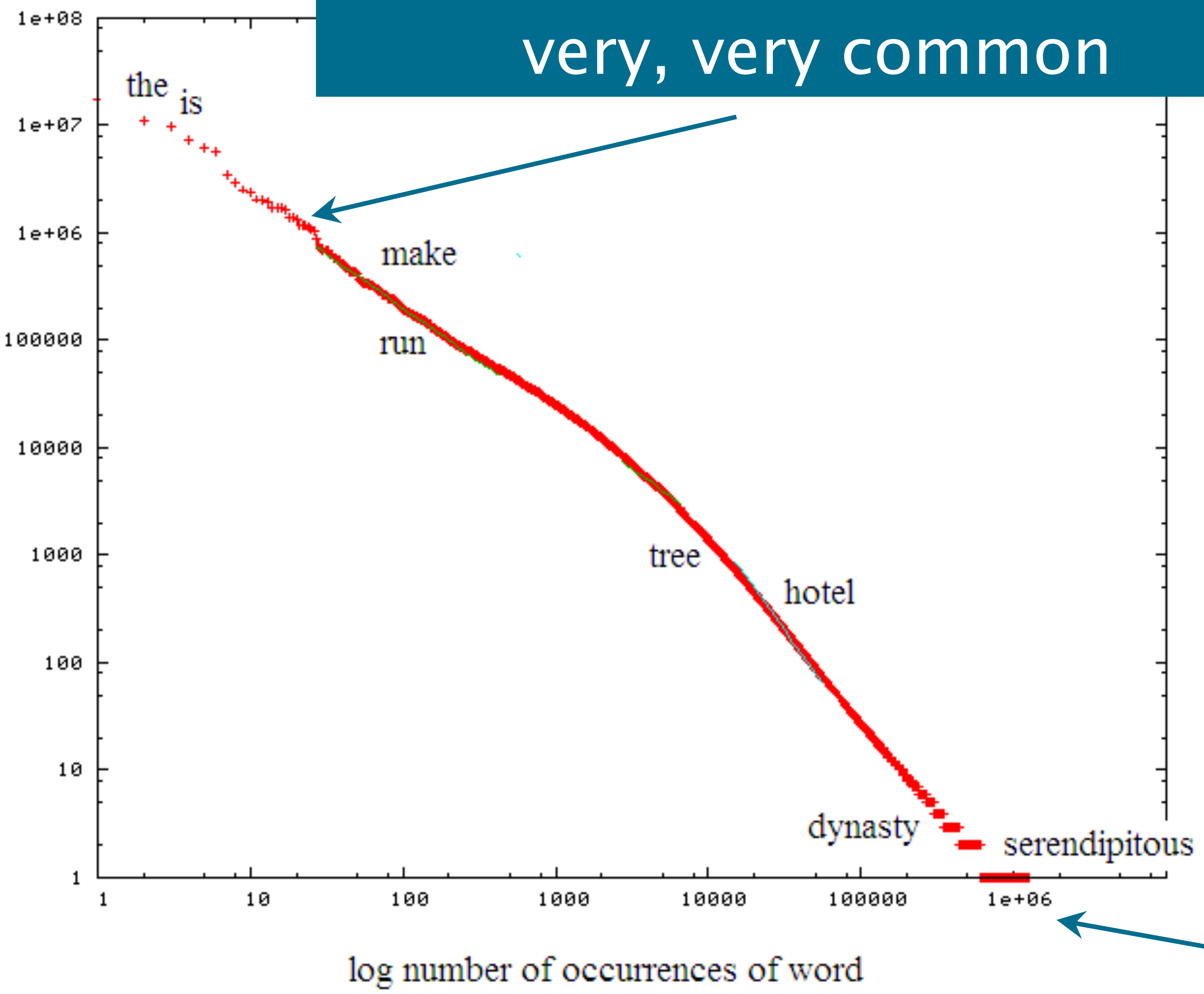
Results: a different corpus



```
plotngrams(tallies)
```

Frequencies always follow the same pattern

Zipf’s Law: word frequencies follow a power-law* distribution function



The most common words are very, very common

The frequency of any word is inversely proportional to its rank in the frequency table

* Occurs when the frequency of some event varies as a power of some attribute of that event: $f(x) = ax^k$

Most words occur only 1 or 2 times

Frequencies always follow the same pattern

In fact, *n*-grams tend to follow Zipf’s law as well!

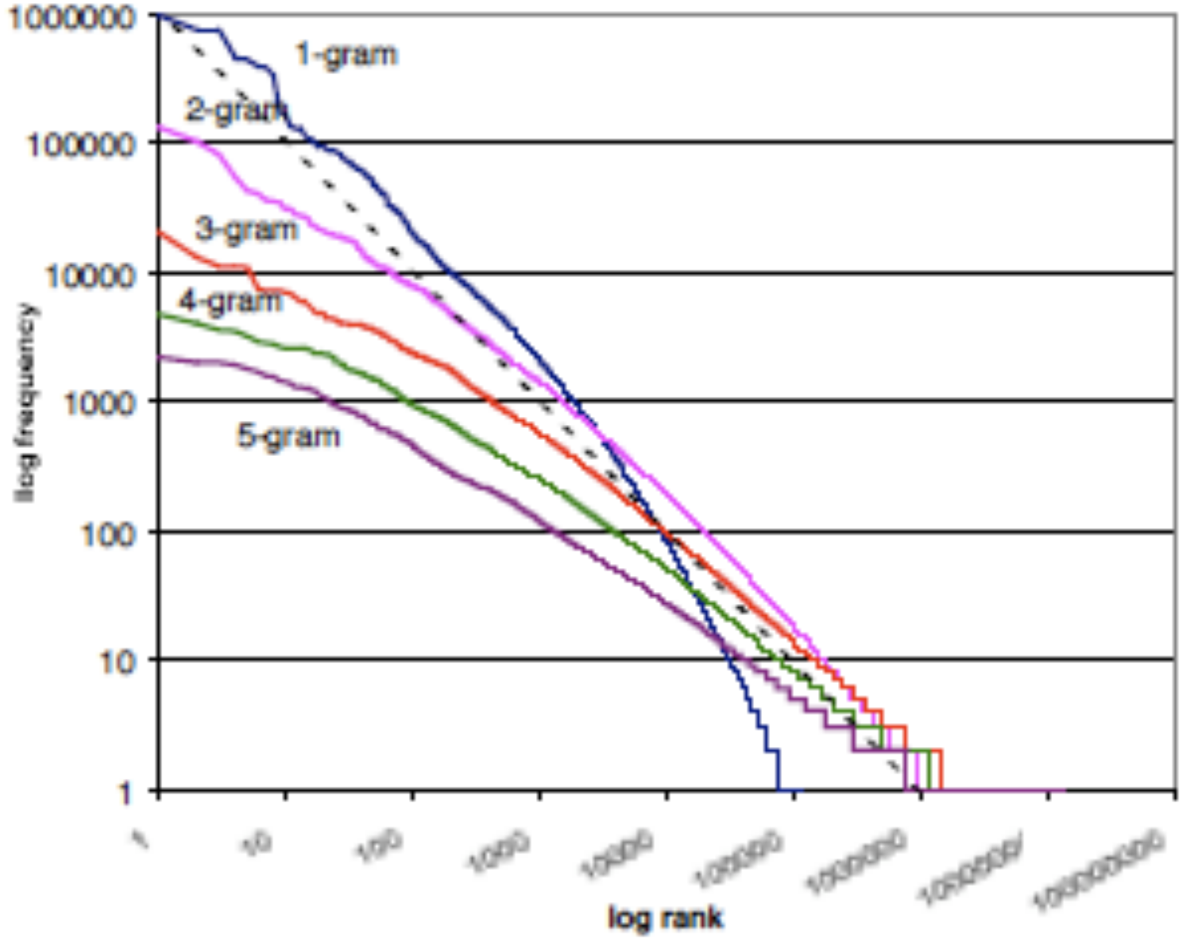


Figure 4 Zipf curves for the WSJ87 corpus

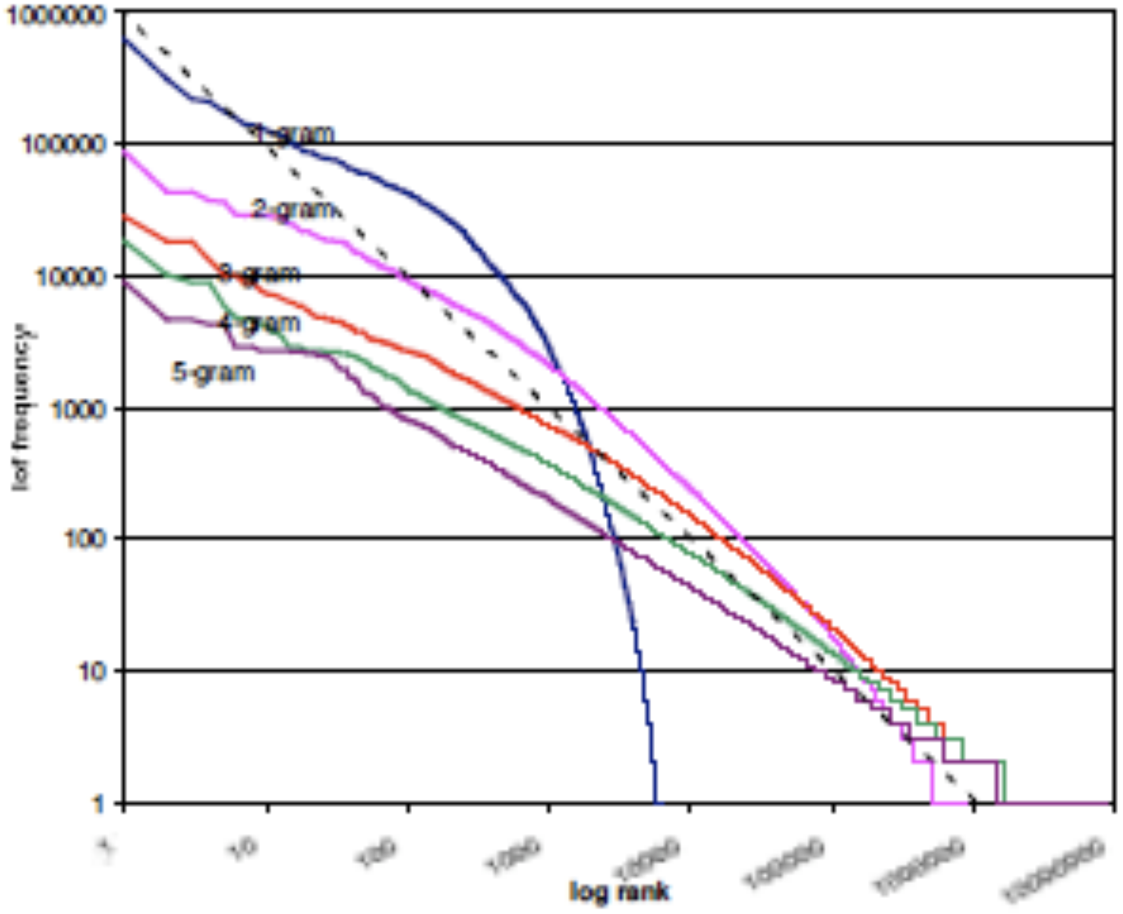


Figure 7 Zipf curves for the TREC Mandarin corpus

Ha et al, 2002

Zipf's law makes life difficult

- Because of it, almost all of our n-grams will be sparsely observed in any given corpus
- Many are ungrammatical and you wouldn't expect to observe them (the of it)
- ... But many are low-frequency but grammatical: probably not going to be observed in any given corpus, but which we want to allow for the possibility of one day seeing (bigoted actuary)
- As a result, the maximum likelihood n-gram model of a corpus predicts that you will not see many things that you actually might
- In essence, it overfits the data (as often happens when we rely on likelihood only, without a prior!)

A consequence of overfitting...

Step 1

Estimate the n-gram probabilities on a training corpus

Step 2

Test to see how accurate it is on a different corpus

Scorching heat to return to southern Australia, total fire bans in place

Updated Tue 28 Jan 2014, 11:48am AEDT

Residents across much of southern Australia are bracing for another heatwave, with temperatures forecast to reach into the 40s in some areas today.

Total fire bans have been issued across South Australia, Victoria and Tasmania ahead of the extreme heat.

Adelaide's maximum temperature today is expected to be 41 degrees Celsius, with 40C on Friday, 41C on Saturday and 40C on Sunday.

A catastrophic fire danger rating has been issued for the state's lower south-east.

PHOTO: Tmperatures forecast to reach into the 40s in some areas today. (ABC News: Amy Simmons)

Heatwave 'one of the most significant' on record, says Bureau of Meteorology

January 20, 2014

Read later

Heat takes its toll on a ballboy at the Australian Open: officials invoked the tennis tournament's "extreme heat policy".

Last week's heatwave that baked most of south-eastern Australia rivalled the intensity of the searing temperatures that preceded the Black Saturday bushfires almost five years ago, according to analysis by the Bureau of Meteorology.

A "dome of very hot air" formed over WA in the second week of January, breaking records in that state before heading eastward, the bureau said in a special climate statement. The warmth has since shifted north to Queensland, forming heatwave conditions over most of that state.

http://www.smh.com.au/environment/weather/heatwave-one-of-the-most-significant-on-record-says-bureau-of-meteorology-20140120-314od.html

A consequence of overfitting...

Step 1

Estimate the *n*-gram probabilities on a training corpus

Residents across much of southern Australia are bracing for another heatwave, with temperatures forecast to reach into the 40s in some areas today. $ Total fire bans have been issued across South Australia, Victoria and Tasmania ahead of the extreme heat. $ Adelaide's maximum temperature today is expected to be 41 degrees Celsius, with 40C on Friday, 41C on Saturday and 40C on Sunday. $ A catastrophic fire danger rating has been issued for the state's lower southeast. $ Country Fire Service state coordinator Brenton Eden says the weather conditions in South Australia could not be worse. $ We are facing a horror day when we already have existing fires burning in the state he said. $ Firefighters have been battling the Bangor fire in the southern Flinders Ranges for a fortnight. $ Victoria is also on fire alert, with temperatures expected to reach 39C in Melbourne and up to 42C in the state's west. $ The Country Fire Authority has listed an extreme fire rating for the South West and Wimmera regions, and says bushfires could become uncontrollable in today's extreme conditions. $ Along with scorching heat, winds of up to 40 kilometres per hour are forecast for western Victoria. $ Those conditions would lead to fires being quite uncontrollable if a fire started, CFA spokesman Steven Walls said. $ Most of Victoria will be subject to potentially very significant fire conditions, so we are asking all Victorians to take particular care when they're outdoors with anything that might cause fires, that includes machinery. $

```
tallies1 <-
getbigramtallies("weather.txt")
```

Step 2

Test to see how accurate it is on a different corpus

Last week's heatwave that baked most of south-eastern Australia rivalled the intensity of the searing temperatures that preceded the Black Saturday bushfires almost five years ago, according to analysis by the Bureau of Meteorology. $ A dome of very hot air formed over WA in the second week of January, breaking records in that state before heading eastward, the bureau said in a special climate statement. $ The warmth has since shifted north to Queensland, forming heatwave conditions over most of that state. $ While the heatwave broke few records for daily maximums SA’s Mt Gambier being one exception many sites set records for prolonged heat. $ For Victoria, Tasmania, southern NSW and the southern half of SA, the heatwave ranked alongside those of January February 2009, January 1939 and January 1908 as one of the most significant on record, the report said. $ Extreme heat persisted for a longer period (last week) than it did in those heatwaves over some areas, the report said. $ These areas included Melbourne and Adelaide, and other coastal regions of Victoria and SA. $ Victoria, for instance, had its hottest four-day period on record for both maximum and average heat. $ Melbourne’s average temperature on Thursday was 35.45 degrees, narrowly eclipsing the previous high of 35.4 set on January 30, 2009. $ The heat took its toll on public health, with Victoria’s ambulance services handling 77 calls on Friday for cardiac arrests, almost six times the number for a typical summer’s day. $ Play was also disrupted in the Australia Open on Thursday, with officials invoking the tennis tournament’s extreme heat policy. $

```
tallies2 <-
getbigramtallies("weather2.txt")
```

A consequence of overfitting...

Question 1

What percent of words (unigrams) occurred in the second corpus that did not occur in the first?

75.4%

Question 2

What percent of bigrams occurred in the second corpus that did not occur in the first?

93.3%

```
overlap <- calculateoverlap(tallies1,tallies2)
```

A consequence of overfitting...

- These are fairly poor because both corpora were *tiny*: for more accurate estimates, you need millions of words
  - Because of Zipf's law, there will be always a lot of low-frequency words or n-grams that only occur once, or never occur but are grammatical
- MLE highly overfits: it doesn't allow for unseen words

A consequence of overfitting...

- These are fairly poor because both corpora were *tiny*: for more accurate estimates, you need millions of words
  - Because of Zipf's law, there will be always a lot of low-frequency words or n-grams that only occur once, or never occur but are grammatical
- MLE highly overfits: it doesn't allow for unseen words

How can we fix this problem?

Additional references (not required)

N-gram models

- Manning, C., & Schutze, H. (1999). Foundations of statistical natural language processing. Chapter 5: 191-203

Zipf's law for phonemes

- Tambovtsev, Y., & Martindale, C. (2007). Phoneme frequencies follow a Yule distribution. *SKASE Journal of Theoretical Linguistics 4*(2): 1-11.

Word segmentation

- Frank, M., Goldwater, S., Griffiths, T., & Tenenbaum, J. (2007). Modeling human performance in statistical word segmentation. *Proceedings of the 29th conference of the Cognitive Science Society.*
- Goldwater, S., Griffiths, T., & Johnson, M. (2009). A Bayesian framework for word segmentation: Exploring the effects of context. *Cognition 112*: 21-54.
- Venkataraman, A. (2001). A statistical model for word discovery in transcribed speech. *Computational Linguistics 27*(3): 351-372.