Should Corpora be Big, Rich, or Dense?

Greg P. Kochanski, Chilin Shih, and Ryan Shosted

The University of Oxford, UK
The University of Illinois, Urbana-Champaign

Outline:
* Zipf's law
* N=1000 and Poisson statistics
* Small, large, and huge corpora
* as corpus size increases, expanding the "magic circle with N>1000"
* Adding data adds questions
* Adding data may cause you to describe a nonexistant average: Glaswegian "wee".
* Identifying a word or a dialect (are two instances equivalent?)
* Idea of data density
* Plots of data density versus corpus size
* Experiments: Sherlock Holmes and avoiding Zipf's law
* Multichannel data
* Trade-off relationships

We start from a simple, clean experiment:

- Reading naturally produced text
- Uniform text style
- Uniform dialect
  - Record your local dialect
  - Carefully designed, reasonably objective criteria
  - Be willing to reject 60% of volunteers
- Uniform recording conditions
  - Same microphone (keep frequency response constant)
  - Same recording booth (keep echoes constant)
  - Unchanged wiring and pre-amp to keep noise constant.
- Uniform reading style
  - Avoid people reading in funny voices
  - Avoid intense boredom

Collecting 100 hours of such data in 6 months is possible.
Total: $2*10^5$ words.

Zipf's Law

The frequency of a word is inversely proportional to its frequency ranking.

In the English language, the probability of encountering the r^{th} most common word is given roughly by

$$P(r)= 0.1/r$$

for r up to 1000 or so.

- For example, the most frequent word (English ”the”) may account for 7% of a corpus. The next frequent word occurs 3.5% of the time.
- This law breaks down for less frequent words, since the harmonic series diverges.

Wolfram MathWorld

http://mathworld.wolfram.com/ZipfsLaw.html

George Zipf: statistician and linguist.

Selected Books:

- *Selected studies of the principle of relative frequencies in language.* *Harvard University Press* (1932).
- *Psycho-biology of Languages.* *Houghton Mifflin* (1935), MIT Press (1965).
- *Human Behavior and the Principle of Least Effort.* *Addison-Wesley* (1949).

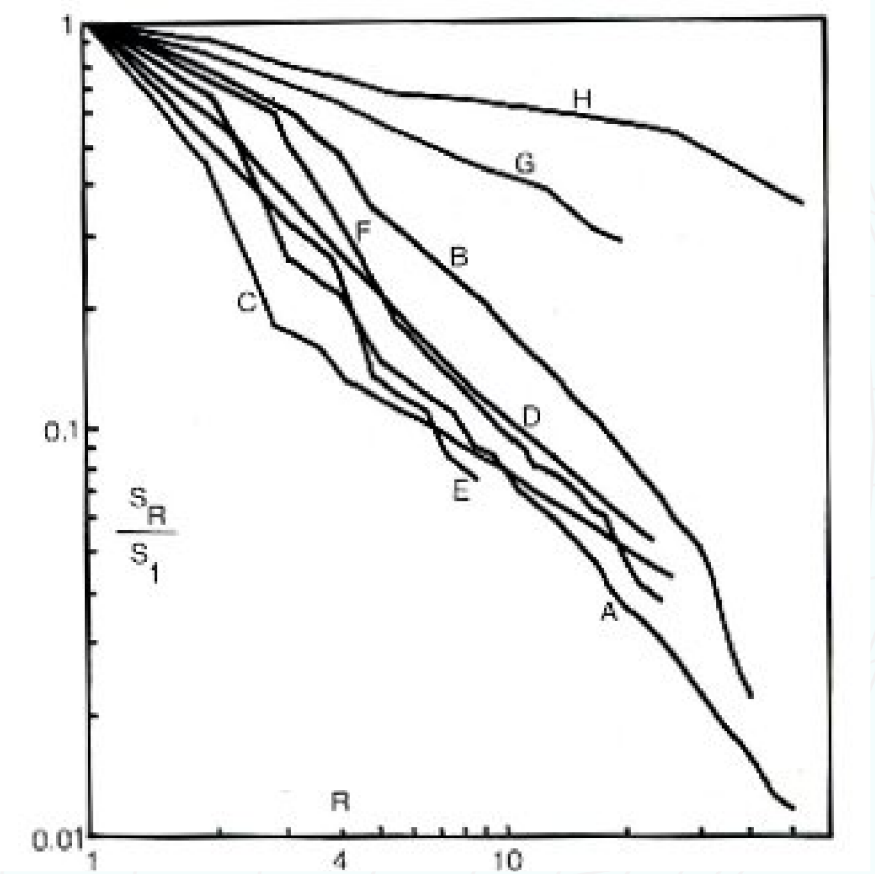
The Implications of Zipf's Law

- There are very few high frequency items.
- Most things in a language are rare.
- There are many, different rare things.

- This is why we need large corpus: to catch the rare things.

Power laws for other kinds of items

- A: Populations of all countries
- B: Number of ships built by all countries
- C: Students at English universities
- D: Building Societies by assets
- E: Populations of World's religions
- F: US insurance companies by staff
- G: World languages
- H: English public schools by students



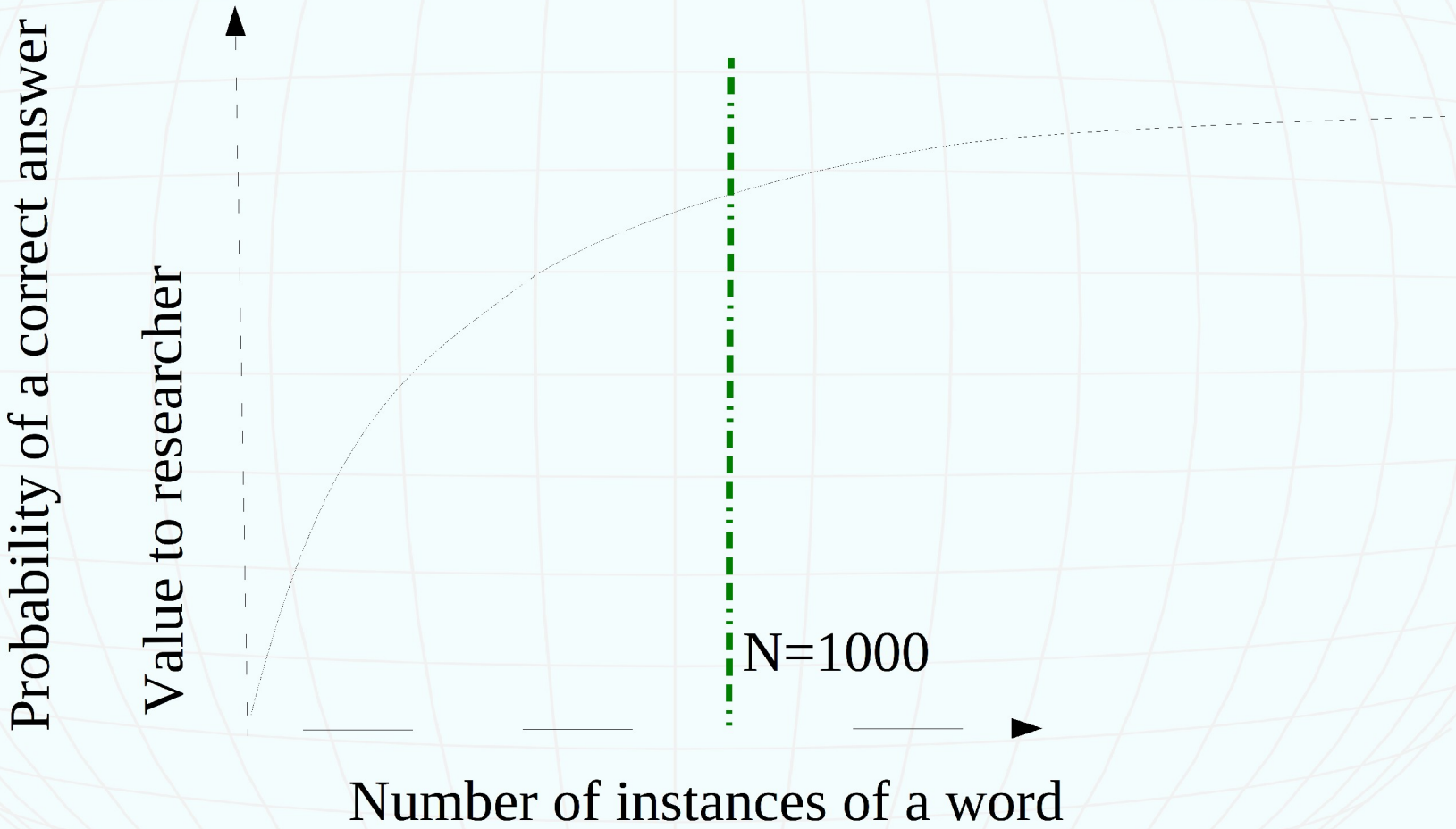
Geoff Kirby (1985). Zipf's Law. UK Journal of Naval Science 10(3) pp 180-185.
http://en.wikipedia.org/wiki/Zipf%27s_law

Zipf's Law is a power law. Power laws are now understood to arise in many statistical situations whenever there is no special frequency (or rank) picked out by nature. The interesting thing about a power law is its slope.

Early Research with Very Large Corpora

• Chen, He-Qin 陳鶴琴 (1892-1982). Psychology and early childhood education.
• Ranked about 5000 most frequent Chinese characters calculated from a corpus of around 900,000 characters in the book *Applied Character List in Vernacular Writing* 語體文應用字匯 .
• Each character from a selected source book was cut and placed into sorting pigeon holes that lined the walls. When the cutting and sorting were done for a book, the frequency count of each character was tallied by counting paper squares in the pidgen hole. When all books were done, compute grand total.
• The project started in 1920 and preliminary results were published in 1922.
• Unfortunately, part of the data was destroyed in a fire in 1923.It took Chen and 9 assistants 2 to 3 years, or 20 to 30 man-years, to complete the work. The revised version was finally published as a book in 1928.

How much data do you need for each word?



Large and Huge Phonetic Corpora

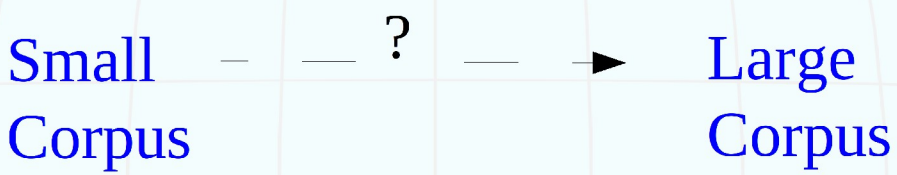
Research on	How big is a "large" corpus?	How large is a "huge" corpus?
Individual phones	> 10^3 words	> 10^5 words
Triphones	> 10^5 words	> 2 . 10^6 words
Triphones with prosody	> 10^6 words	> 4.10^9 words
Individual words	> 3 x 10^5 words	> 10^9 words
Word bi-grams	> 10^7 words	> 10^{15} words

1000 or more instances of the most frequent objects

1000 or more instances of most objects

What Changes?

Uniform scripted reading experiment



What Changes?

Uniform scripted reading experiment

Small Corpus — ? → Large Corpus

How to enlarge it?

- Bring in different dialects.
- Broader range of material to read.
- Outsource the recruitment.
- Outsource the recording process.
- Use unscripted speech.

Questions that come with enlargement.

Uniform scripted reading experiment

Small Corpus — ? → Large Corpus

Do these two people have the same dialect?

What dialect is the best match?

What was the content?

What is the pattern of room echoes?

How formal were the recording conditions?

How much enthusiasm in the reading?

How to enlarge it?

- Bring in different dialects.
- Broader range of material to read.
- Outsource the recruitment.
- Outsource the recording process.
- Unscripted speech.

What to do about those questions?

- Select for uniformity
- Ask broader questions
- Build a model that unifies all your data

What to do about those questions?

Select for uniformity?

Extreme example:

- A corpus of **Southern British English (SBE)**, for studies of intonation, Enlarged by adding **Singapore English** and **Nigerian English**.

- **Singapore English** has substantial differences from **SBE**: unstressed multisyllabic words (Deterding 1994), and rhythmic differences (Low, Grabe, and Nolan 2001).

- There is evidence that **Nigerian English** is better described as a tone language, with a specification for each syllable (Gussenhoven 2010). **SBE** is not.

Adding these extra dialects tells you virtually nothing new about SBE prosody.

David Deterding, "The intonation of Singapore English", *Journal of the International Phonetic Association* (1994), 24: 61-72
doi:10.1017/S0025100300005077
Low, E.L., Grabe, E. and Nolan, F. (2001). Quantitative characterisations of speech rhythm: Syllable-timing in Singapore English. Language and Speech 43 (4), 377-401.
Gussenhoven, C and Udofot, I., "Word melodies vs. pitch accents: A perceptual evaluation of terracing contours in British and Nigerian English", Proceedings of Speech Prosody, 2010.

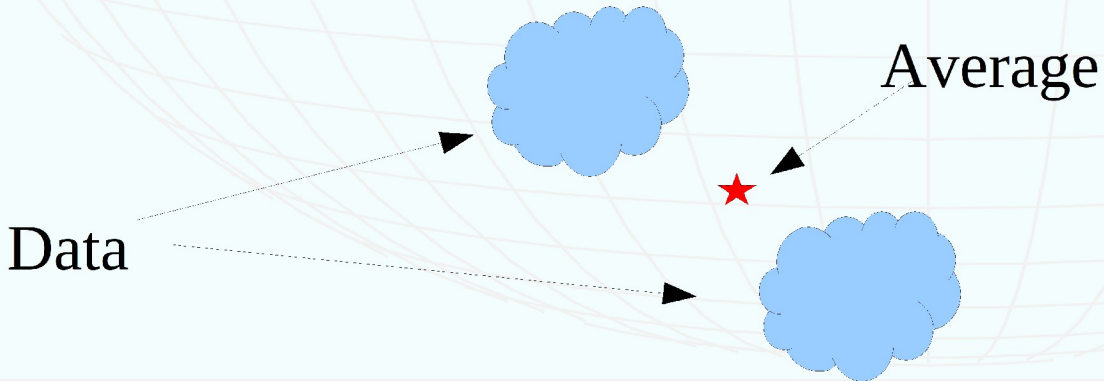
What to do about those questions?

Ask broader questions?

An extreme example:
Expand a Southern British English corpus to include Glaswegian.
Can we broaden our research questions to speak of “English” rather than SBE?

“Wee” (Glasgow) = “Little” (SBE).

• It would be silly to average the acoustic properties of “the word meaning 'small".”



This is a case of taking the average of a bimodal distribution. An average is an artificial quantity: it can appear in places where data doesn't exist.

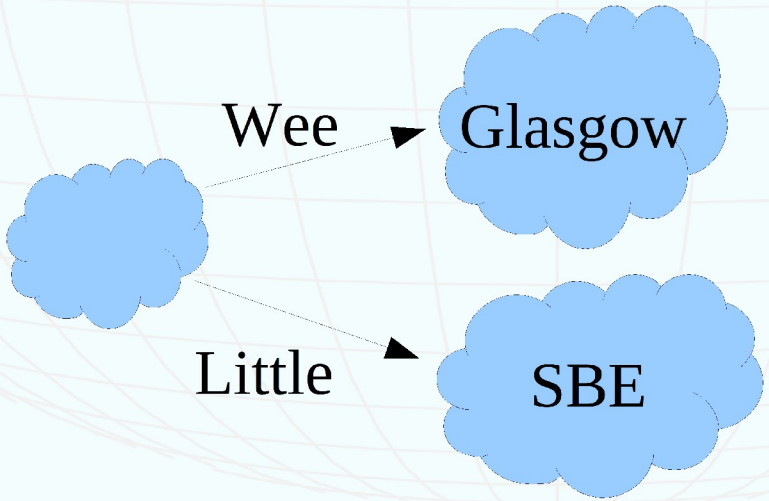
What to do about those questions?

Ask broader questions?

An extreme example:
Expand a Southern British English corpus to include Glasgow.

“Wee” (Glasgow) = “Little” (SBE).

• We might find that people typically use “wee” when talking to a Glaswegian, but use “little” when talking to a SBE speaker.
• We might interpret this as evidence for strong lexical adaption.



What to do about those questions?

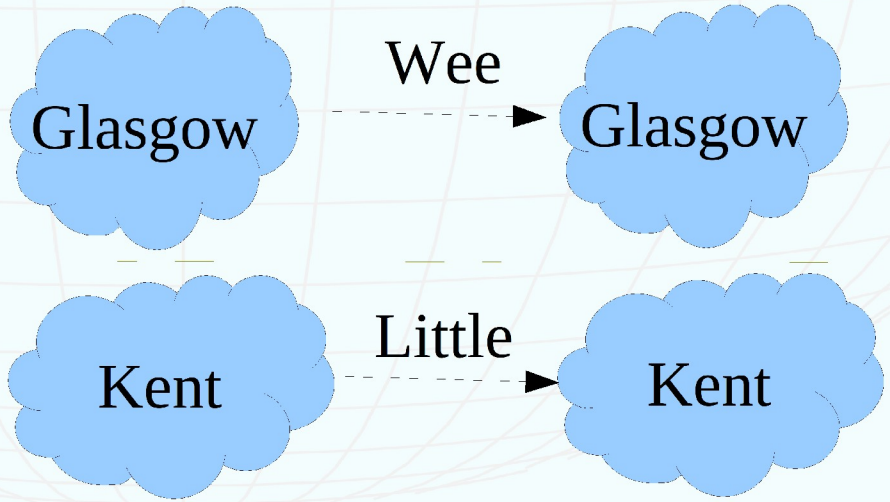
Ask broader questions?

An extreme example:
Expand a Southern British English corpus to include Glasgow.

“Wee” (Glasgow) = “Little” (SBE).

- We might find that people typically use “wee” when talking to a Glaswegian, but use “little” when talking to a SBE speaker.
- We might interpret this as evidence for strong lexical adaption.
- But really, it means that SBE people talk to SBE people and Glaswegians talk mostly to Glaswegians.

Six hours of travel.



What to do about those questions?

Build a model that unifies all your data

This can be complicated,
and it means you need to answer the questions.

What to do about those questions?

Select for uniformity

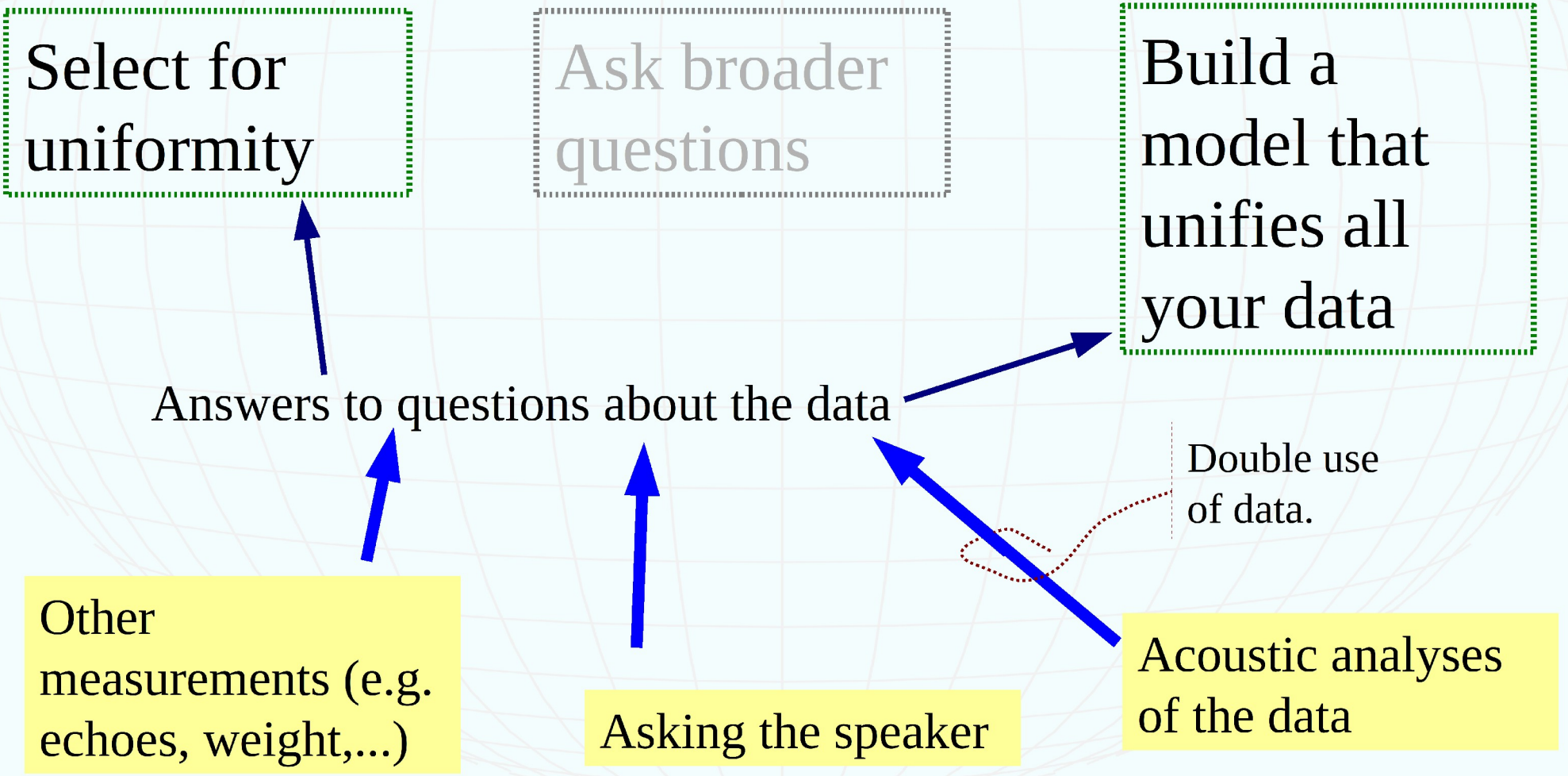
Ask broader questions

Build a model that unifies all your data

You still have to attempt to answer them.

Entropy of the acoustic properties
Uncontrolled variation in acoustic properties
Uncontrolled variation
Variability
Reading words - lab speech
Reading natural texts
Unscripted speech
Naturalness
Unpredictable variation in the transcript
Entropy of the phonological transcription

What to do about those questions?



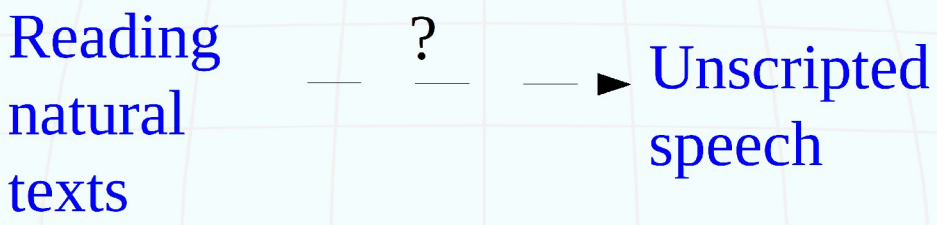
Using data twice

One should not use your data twice in pursuit of one conclusion.

Why?

- You're only getting information when there is a chance that you may be surprised.
  - Man bites dog is news; dog bites man is not.
- ...and there's less (or no) surprise the second time.
- ...so you have less information than you think,
- ...so your conclusions are not as reliable as you think.

Double counting data



Question: *What's that phone?*

Let's look at a real (though exaggerated) case.

louie-louie

Chorus: Oh, Louie, Louie, Oh, No,
Get her way down low.

Oh, Louie, Louie, Oh, Baby,
Get her down low.

A fine little girl awaiting for me
she's just a girl across the way
Well I'll take her and park all alone
She's never a girl I'd lay at home.

(Chorus Repeat)

LOUIE LOUIE

Fine little girl waits for me get your thrills
across the way girl I dream about is all alone
She never could get away from home

Lou-Ai Lou-Ai Oh, no. Grab her way down low.
(Repeat) This line least clear.

There is a fine little girl waiting for me.
She is just a girl across the way.
Then I take her all alone,
She's never the girl I lay at home..

"LOUIE - LOUIE"

"LOUIE, LOUIE...OH YEA, A-WAY WE GO
YEA, YEA, YEA, YEA. YEA
LOUIE, LOUIE...OH BABY, A-WAY WE GO

"A FINE LITTLE GIRL - SHE WAIT FOR ME
ME CATCH THE SHIP - A-CROSS THE SEA
I SAILED THE SHIP - ALL A-LONE
I NEVER THINK - I'LL MAKE IT HOME

TO: DIRECTOR, FBI (145-2961)

FROM: SAC, DETROIT (145-420)(P)

UNSUB; 45 r.p.m.
Recording "Louie Louie"
POSSIBLE ITOM
(OO: DETROIT)

Re FBI Laboratory letter to Detroit dated 5/17/65, and Detroit letter to Bureau dated 4/22/65.

For the information of the New York Office, the FBI Laboratory advised Detroit by referenced letter that the Department of Justice has previously received a copy of the record, "Louie Louie" from [redacted] with a request that it be reviewed to detemmine if it was an obscene matter. The Department advised that they were unable to interpret any of the wording in the record and, therefore, could not make a decision concerning the matter. Also, that the AUSA at Tampa, Fla. and Hammond, Indiana, have declined prosecution.

If “Louie Louie” were unscripted, phonetics would be impossible.

You cannot measure phonetic properties of X unless you know when you have an instance of X.

--or--

You have a certain amount of information, and you use up some of it to find out what the words or phones are.

How much does it take?

16 bits per word (if you have only a word list),
6 bits/word if you have grammar and semantics.

--or--

5 bits per phone (if you only have an inventory),
1 bit per phone (if you know the phonotactics).

How much do you start with?

Corpus design is important

- Studying dialect phonetics if dialect is defined phonetically.
- Studying dialect phonetics if dialect is defined lexically. ✓
- Studying phonotactics in unscripted corpora.
- Studying phonetics if you have an independent transcription. ✓
- Studying phoneme variability if you don't control the microphone.

Tricky...

- Studying anything in “Louie Louie”

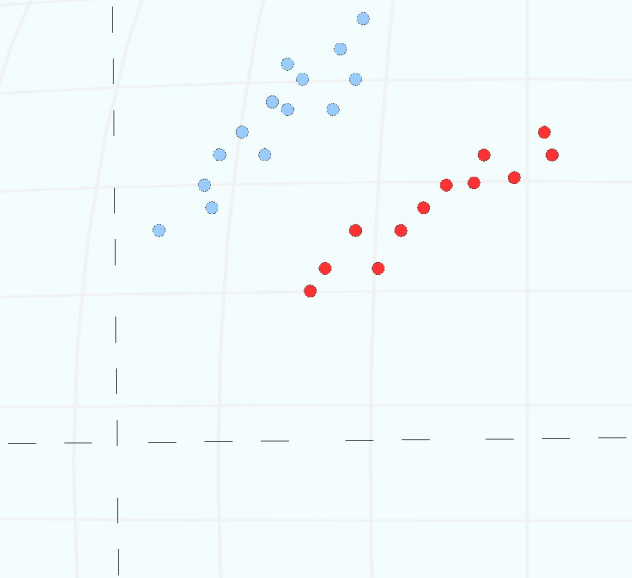
… not just size.

Doing Prosody with Large Corpora

- Prosody is (nearly) unscriptable.
- Human-human agreement in prosodic annotation isn't high.
- Phonology of prosody is unclear.
- Prosody is always in the “Louie Louie” situation.

- There seem to be trade-off relationships in prosody.
  - More than one combination of properties with same function.
  - Loudness, duration, F_0, vowel centrality, articulation speed...

Trade-off relationships

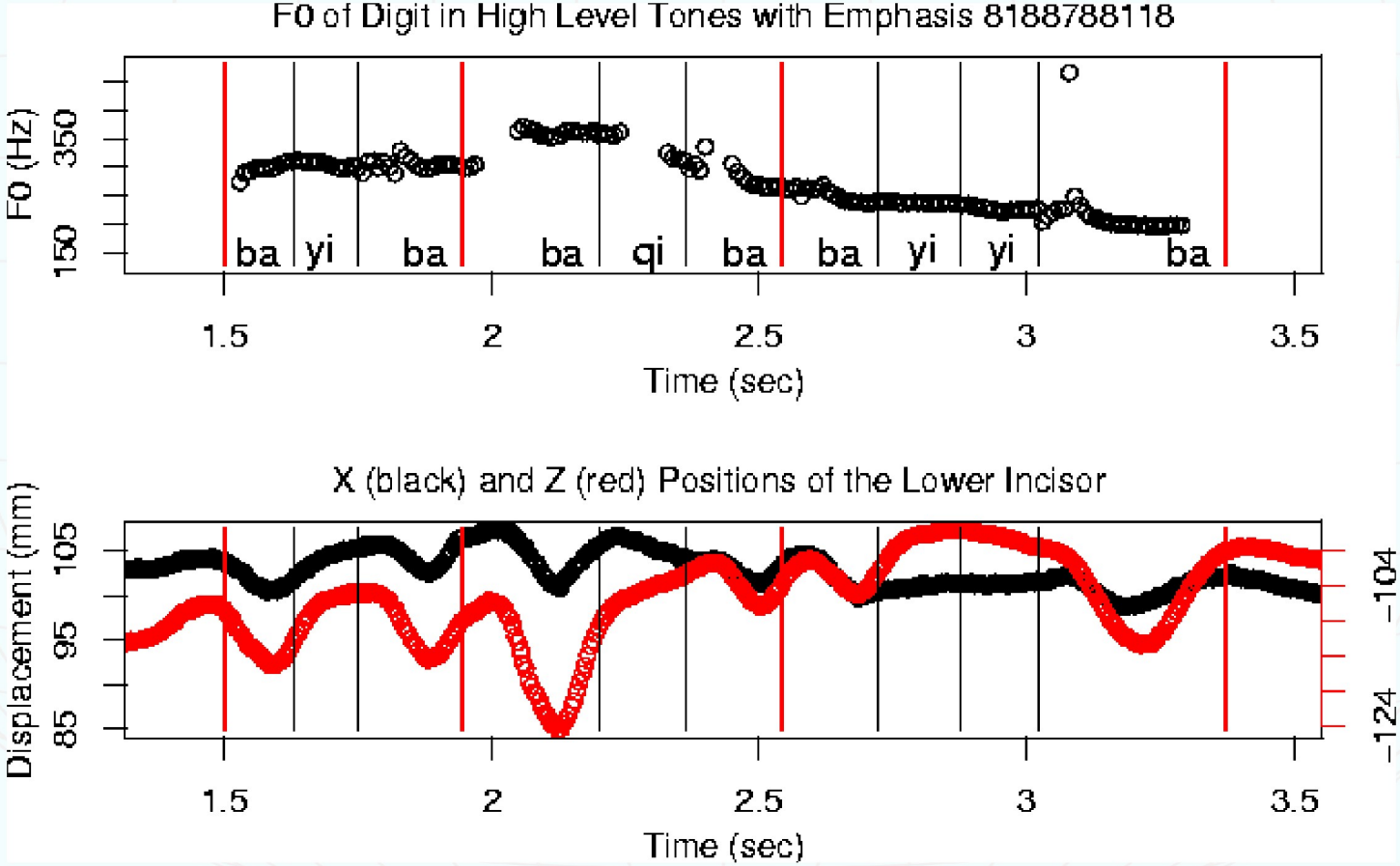


b

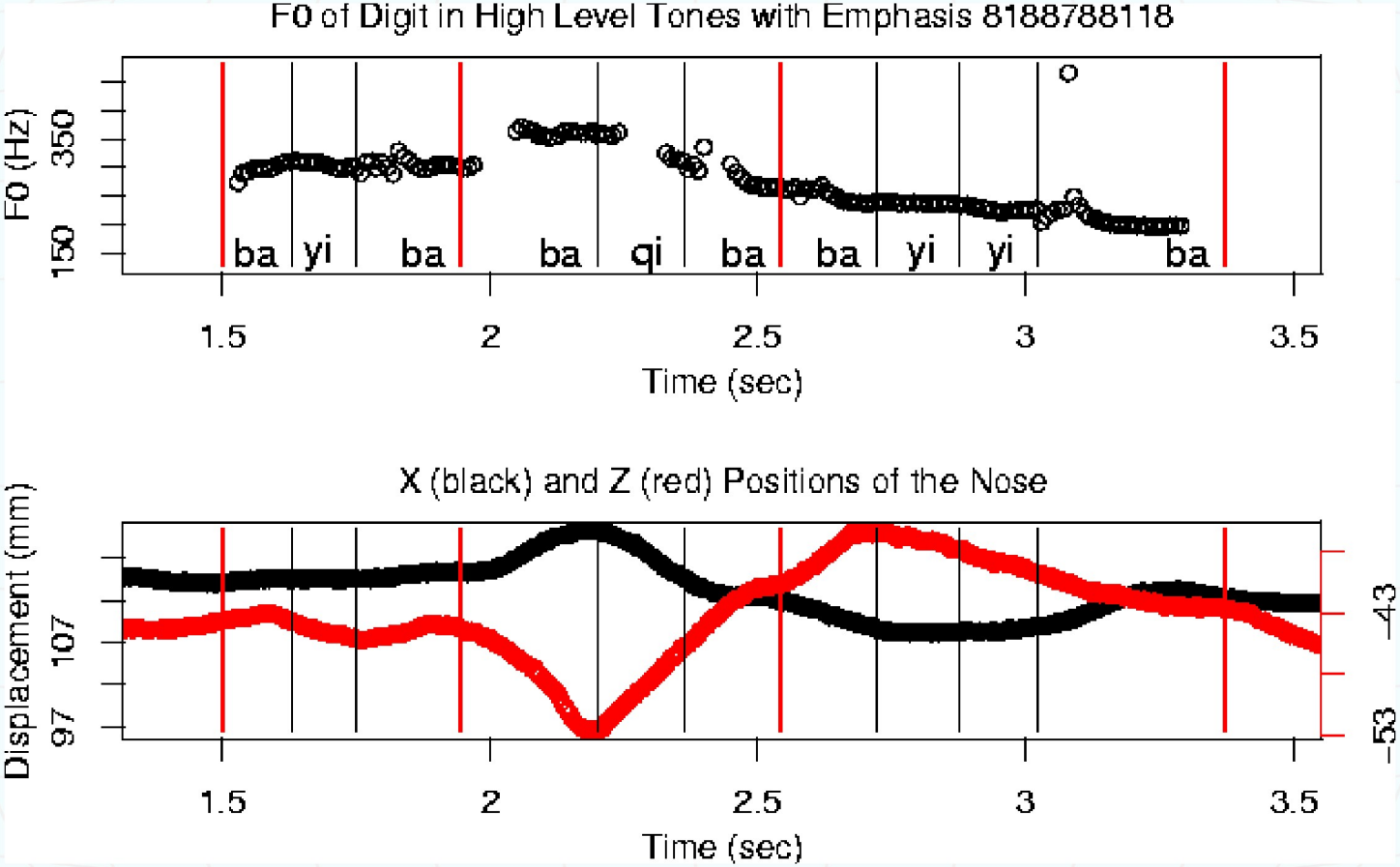
Reading Styles

- Statement
  - (My phone number is) 818-878-8118
- Statement with narrow focus
  - (Is your number 818-578-8118?)
  - (No! My number is) 818-878-8118
- Question
  - (Is your number) 818-878-8118?
- Question with narrow focus
  - (Is your number) 818-878-8118?

Jaw movement correlates with emphasis: Everything being equal, emphasizing a low vowel results in lower jaw position.

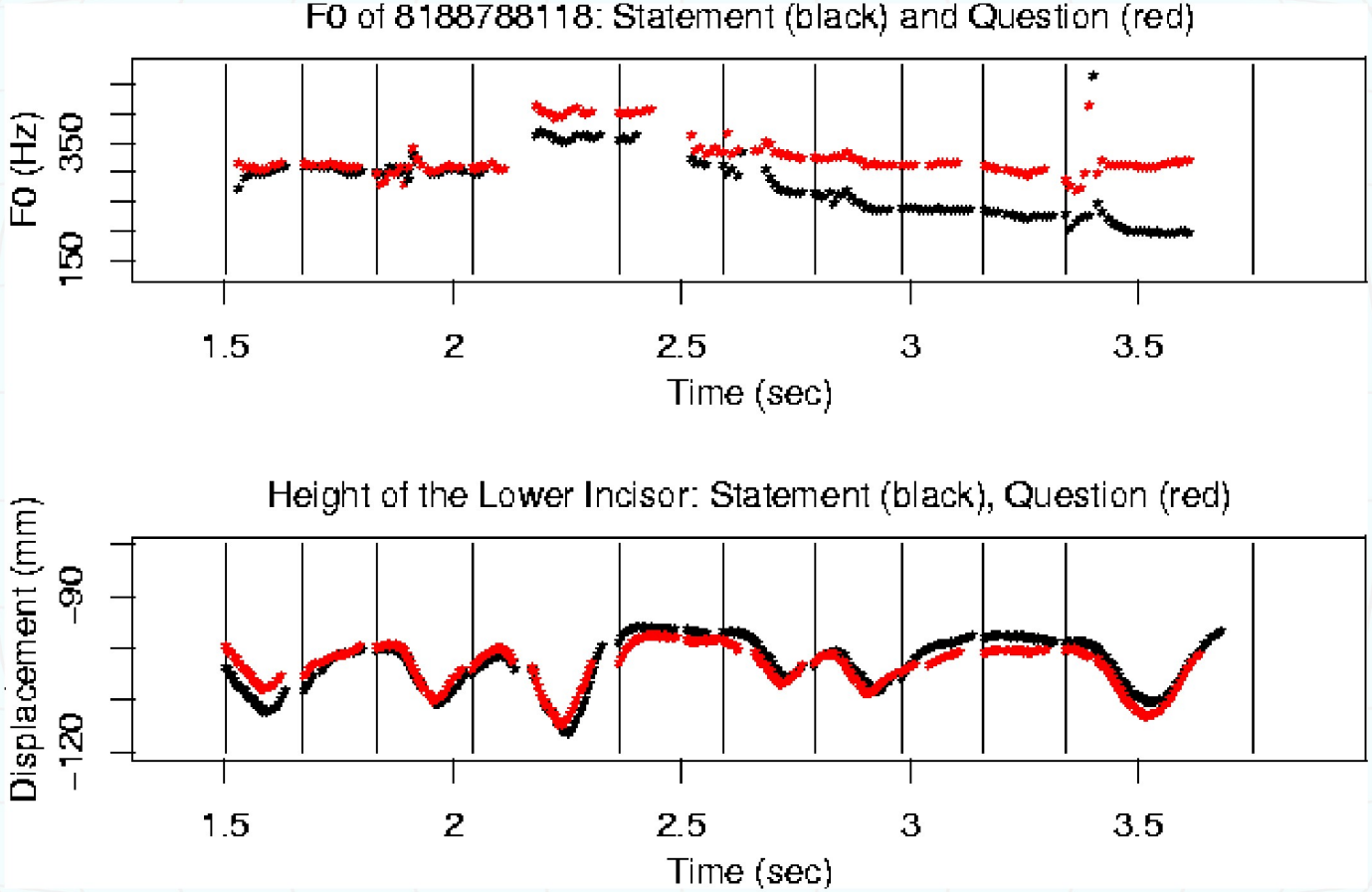


Nose movement correlates with emphasis



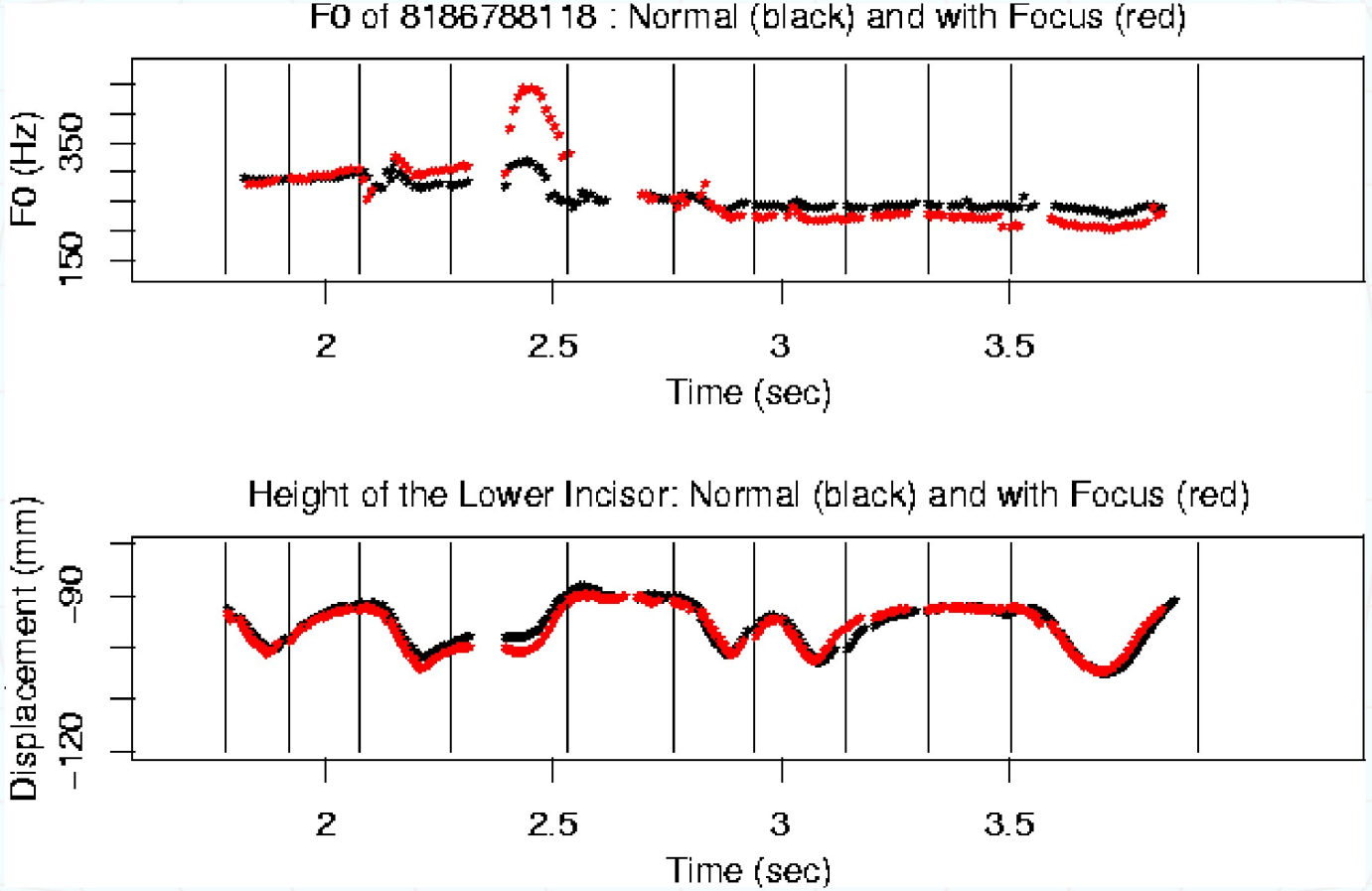
Some of “speech” is visual: shrugs, nods,looking

Question and statement in a sentence with high level tones



F0 tracks are different for ques and statement.
Jaw movements are similar—no diff in emphasis

Normal sentence and the sentence with one digit in narrow focus



Jaw movement shows the effect of emphasis. The magnitude is weak on a glide+mid vowel

Conclusion

- Big, haphazard corpora do some jobs badly.
- Corpus design matters
- You cannot always broaden your question to match your corpus.
- Adding articulatory data helps
  - For prosody
  - For trade-off relationships

Should Corpora be Big, Rich, or Dense?

Greg P. Kochanski, Chilin Shih, and Ryan Shosted

The University of Oxford, UK
The University of Illinois, Urbana-Champaign

Outline:
* Zipf's law
* N=1000 and Poisson statistics
* Small, large, and huge corpora
* as corpus size increases, expanding the "magic circle with N>1000"
* Adding data adds questions
* Adding data may cause you to describe a nonexistant average: Glaswegian "wee".
* Identifying a word or a dialect (are two instances equivalent?)
* Idea of data density
* Plots of data density versus corpus size
* Experiments: Sherlock Holmes and avoiding Zipf's law
* Multichannel data
* Trade-off relationships

We start from a simple, clean experiment:

- Reading naturally produced text
- Uniform text style
- Uniform dialect
  - Record your local dialect
  - Carefully designed, reasonably objective criteria
  - Be willing to reject 60% of volunteers
- Uniform recording conditions
  - Same microphone (keep frequency response constant)
  - Same recording booth (keep echoes constant)
  - Unchanged wiring and pre-amp to keep noise constant.
- Uniform reading style
  - Avoid people reading in funny voices
  - Avoid intense boredom

Collecting 100 hours of such data in 6 months is possible.
Total: $2*10^5$ words.

Zipf's Law

The frequency of a word is inversely proportional to its frequency ranking.

In the English language, the probability of encountering the r^{th} most common word is given roughly by

$$P(r) = 0.1/r$$

for r up to 1000 or so.

- For example, the most frequent word (English "the") may account for 7% of a corpus. The next frequent word occurs 3.5% of the time.
- This law breaks down for less frequent words, since the harmonic series diverges.

Wolfram MathWorld

http://mathworld.wolfram.com/ZipfsLaw.html

George Zipf: statistician and linguist.

Selected Books:

- *Selected studies of the principle of relative frequencies in language.* *Harvard University Press* (1932).
- *Psycho-biology of Languages.* *Houghton Mifflin* (1935), MIT Press (1965).
- *Human Behavior and the Principle of Least Effort.* *Addison-Wesley* (1949).

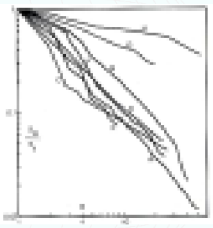
The Implications of Zipf's Law

- There are very few high frequency items.
- Most things in a language are rare.
- There are many, different rare things.

- This is why we need large corpus: to catch the rare things.

Power laws for other kinds of items

- A: Populations of all countries
- B: Number of ships built by all countries
- C: Students at English universities
- D: Building Societies by assets
- E: Populations of World's religions
- F: US insurance companies by staff
- G: World languages
- H: English public schools by students



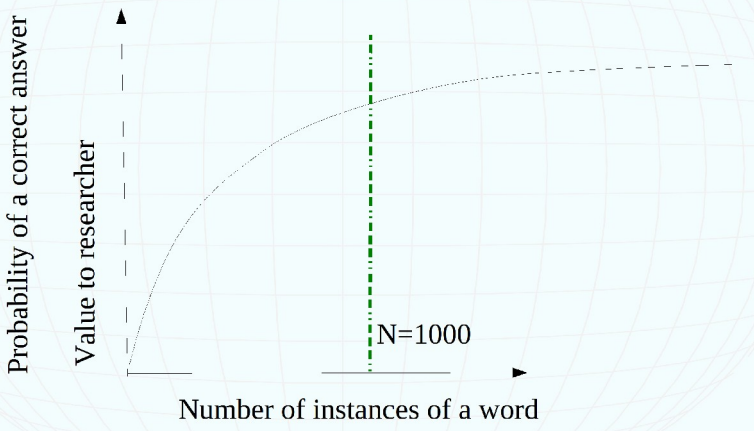
Geoff Kirby (1985). Zipf's Law. UK Journal of Naval Science 10(3) pp 180-185.
http://en.wikipedia.org/wiki/Zipf%27s_law

Zipf's Law is a power law. Power laws are now understood to arise in many statistical situations whenever there is no special frequency (or rank) picked out by nature. The interesting thing about a power law is its slope.

Early Research with Very Large Corpora

• Chen, He-Qin 陳鶴琴 (1892-1982). Psychology and early childhood education.
• Ranked about 5000 most frequent Chinese characters calculated from a corpus of around 900,000 characters in the book *Applied Character List in Vernacular Writing* 語體文應用字匯 .
• Each character from a selected source book was cut and placed into sorting pigeon holes that lined the walls. When the cutting and sorting were done for a book, the frequency count of each character was tallied by counting paper squares in the pidgen hole. When all books were done, compute grand total.
• The project started in 1920 and preliminary results were published in 1922.
• Unfortunately, part of the data was destroyed in a fire in 1923.It took Chen and 9 assistants 2 to 3 years, or 20 to 30 man-years, to complete the work. The revised version was finally published as a book in 1928.

How much data do you need for each word?



Large and Huge Phonetic Corpora

Research on	How big is a "large" corpus?	How large is a "huge" corpus?
Individual phones	> 10^3 words	> 10^6 words
Triphones	> 10^5 words	> 2 . 10^6 words
Triphones with prosody	> 10^8 words	> 4.10^9 words
Individual words	> 3 x 10^5 words	> 10^8 words
Word bi-grams	> 10^7 words	> 10^{15} words

1000 or more instances of the most frequent objects

1000 or more instances of most objects

What Changes?

Uniform scripted reading experiment

Small Corpus — ?—— ► Large Corpus

What Changes?

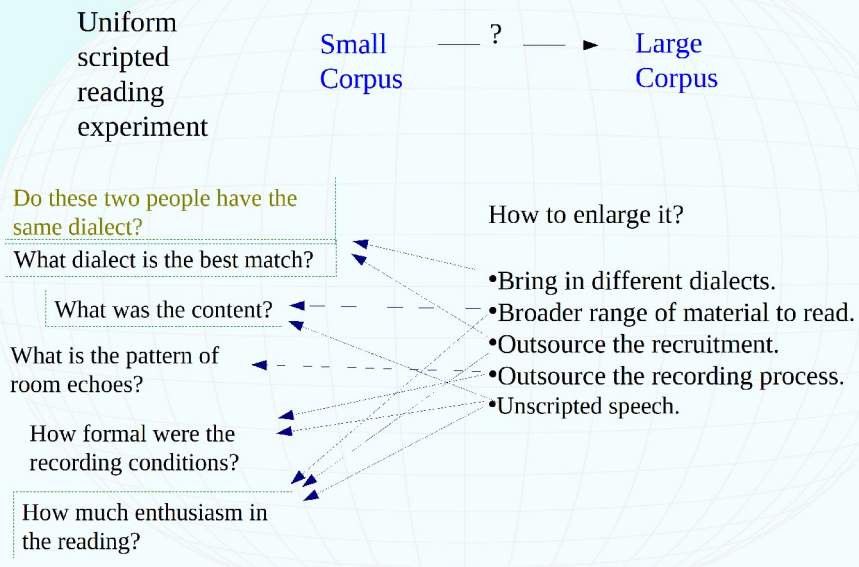
Uniform scripted reading experiment

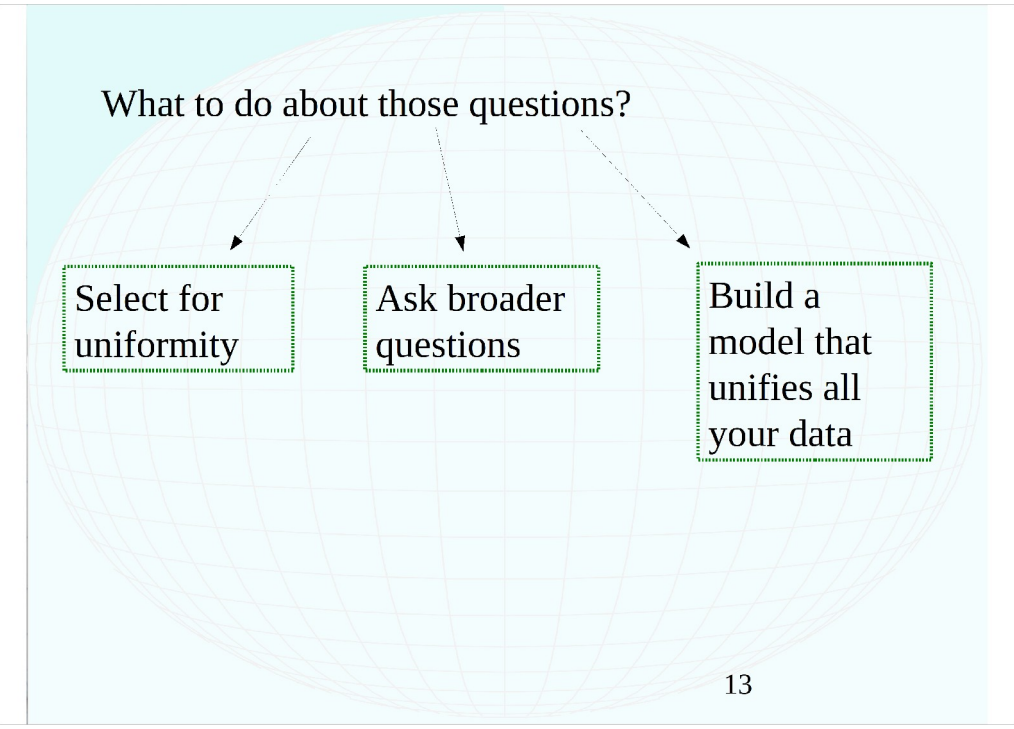
Small Corpus — ?—— ► Large Corpus

How to enlarge it?

- Bring in different dialects.
- Broader range of material to read.
- Outsource the recruitment.
- Outsource the recording process.
- Use unscripted speech.

Questions that come with enlargement.




What to do about those questions?
Select for uniformity
Ask broader questions
Build a model that unifies all your data

What to do about those questions?

Select for uniformity?

Extreme example:

• A corpus of **Southern British English (SBE)**, for studies of intonation, Enlarged by adding **Singapore English** and **Nigerian English**.

• **Singapore English** has substantial differences from **SBE**: unstressed multisyllabic words (Deterding 1994), and rhythmic differences (Low, Grabe, and Nolan 2001).

• There is evidence that **Nigerian English** is better described as a tone language, with a specification for each syllable (Gussenhoven 2010). **SBE** is not.

Adding these extra dialects tells you virtually nothing new about SBE prosody.

David Deterding, "The intonation of Singapore English", *Journal of the International Phonetic Association* (1994), 24: 61-72
doi:10.1017/S0025100300005077
Low, E.L., Grabe, E. and Nolan, F. (2001). Quantitative characterisations of speech rhythm: Syllable-timing in Singapore English. Language and Speech 43 (4), 377-401.
Gussenhoven, C and Udofot, I., "Word melodies vs. pitch accents: A perceptual evaluation of terracing contours in British and Nigerian English", Proceedings of Speech Prosody, 2010.

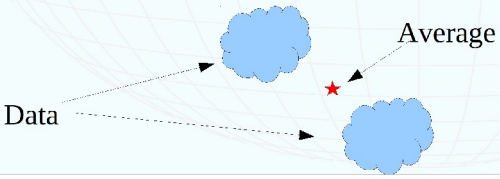
What to do about those questions?

Ask broader questions?

An extreme example:
Expand a Southern British English corpus to include Glaswegian.
Can we broaden our research questions to speak of “English” rather than SBE?

“Wee” (Glasgow) = “Little” (SBE).

• It would be silly to average the acoustic properties of “the word meaning 'small".”



This is a case of taking the average of a bimodal distribution. An average is an artificial quantity: it can appear in places where data doesn't exist.

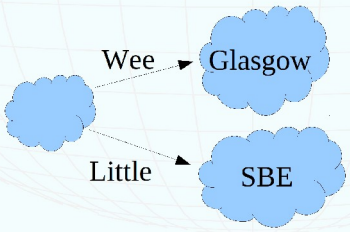
What to do about those questions?

Ask broader questions?

An extreme example:
Expand a Southern British English corpus to include Glasgow.

"Wee" (Glasgow) = "Little" (SBE).

• We might find that people typically use "wee" when talking to a Glaswegian, but use "little" when talking to a SBE speaker.
• We might interpret this as evidence for strong lexical adaption.



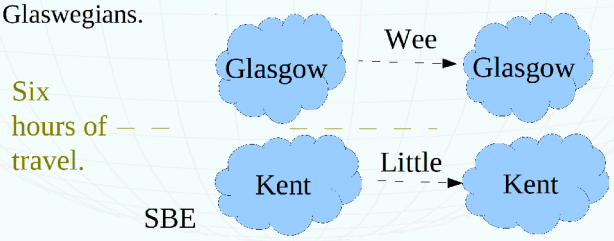
What to do about those questions?

Ask broader questions?

An extreme example:
Expand a Southern British English corpus to include Glasgow.

"Wee" (Glasgow) = "Little" (SBE).

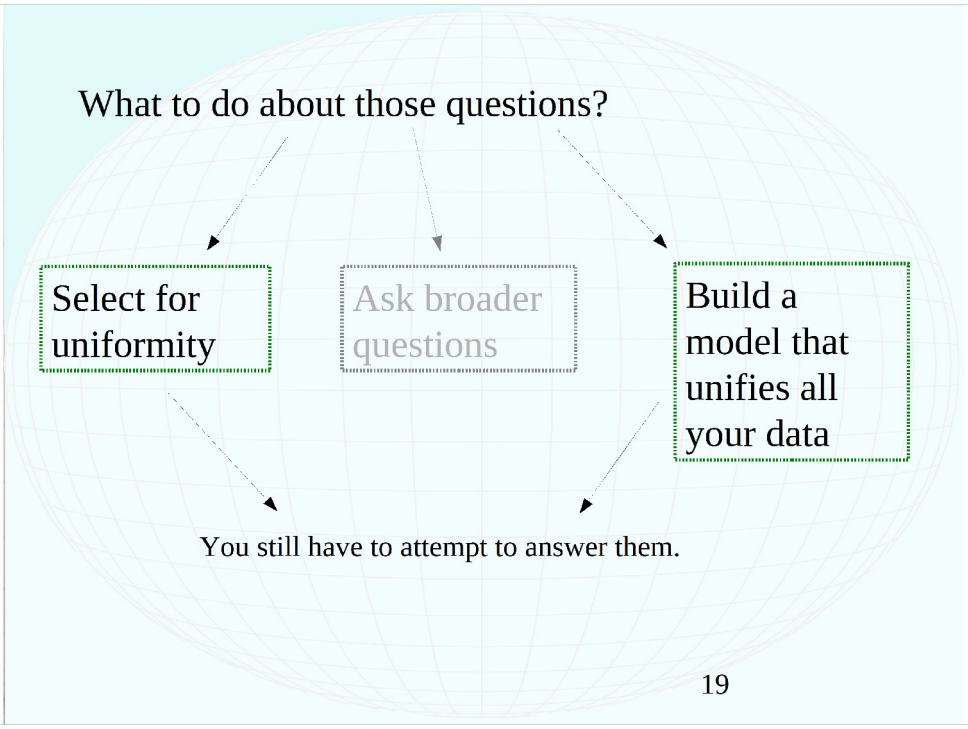
• We might find that people typically use "wee" when talking to a Glaswegian, but use "little" when talking to a SBE speaker.
• We might interpret this as evidence for strong lexical adaption.
• But really, it means that SBE people talk to SBE people and Glaswegians talk mostly to Glaswegians.

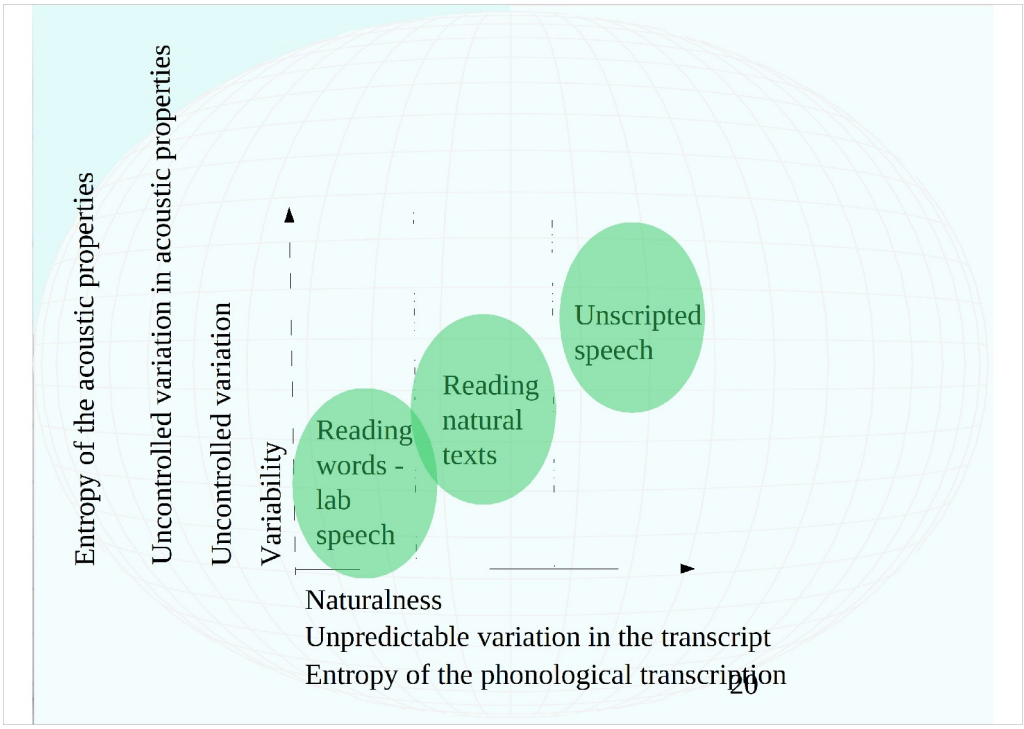


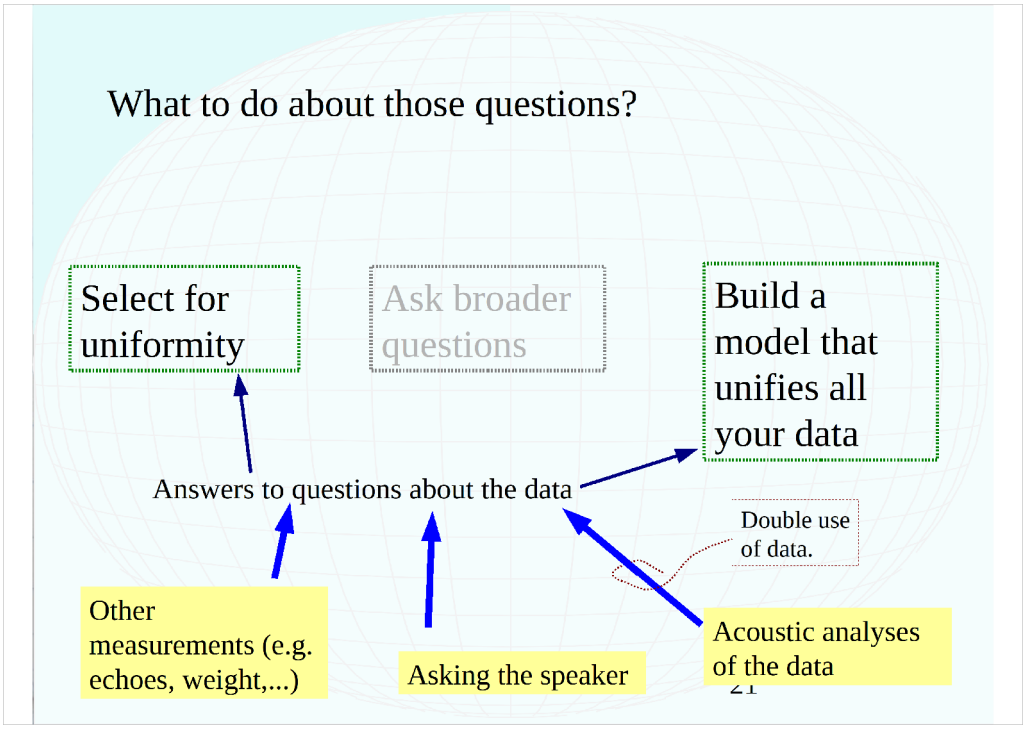
What to do about those questions?

Build a model that unifies all your data

This can be complicated,
and it means you need to answer the questions.


What to do about those questions?
Select for uniformity
Ask broader questions
Build a model that unifies all your data
You still have to attempt to answer them.


Entropy of the acoustic properties
Uncontrolled variation in acoustic properties
Uncontrolled variation
Variability
Reading words - lab speech
Reading natural texts
Unscripted speech
Naturalness
Unpredictable variation in the transcript
Entropy of the phonological transcription


What to do about those questions?
Select for uniformity
Ask broader questions
Build a model that unifies all your data
Answers to questions about the data
Double use of data.
Other measurements (e.g. echoes, weight,...)
Asking the speaker
Acoustic analyses of the data

Double counting data

Reading natural texts ____? ► Unscripted speech

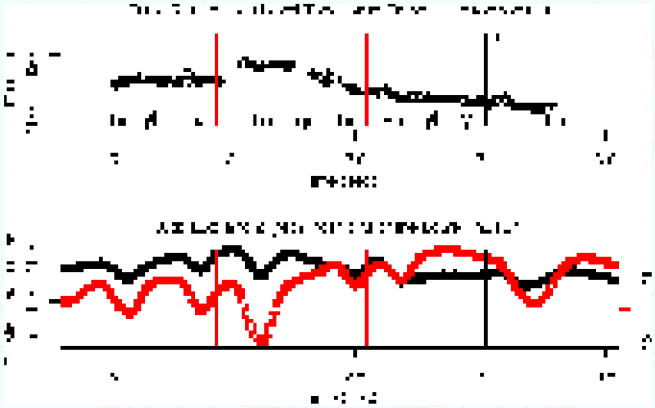
Question: *What's that phone?*

Let's look at a real (though exaggerated) case.

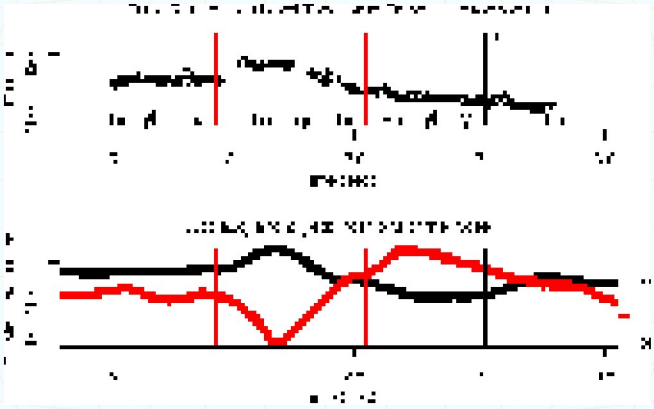
Reading Styles

- Statement
  - (My phone number is) 818-878-8118
- Statement with narrow focus
  - (Is your number 818-578-8118?)
  - (No! My number is) 818-878-8118
- Question
  - (Is your number) 818-878-8118?
- Question with narrow focus
  - (Is your number) 818-878-8118?

Jaw movement correlates with emphasis: Everything being equal, emphasizing a low vowel results in lower jaw position.

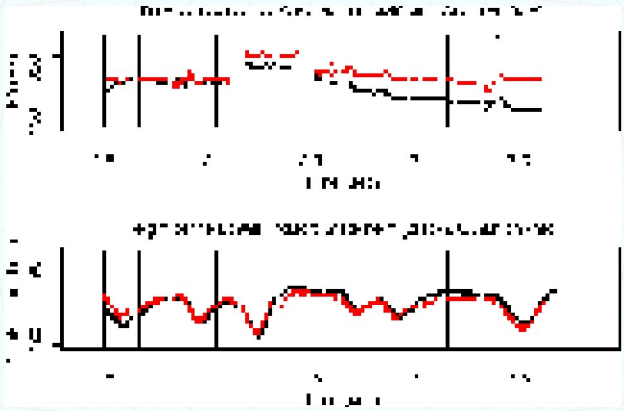


Nose movement correlates with emphasis



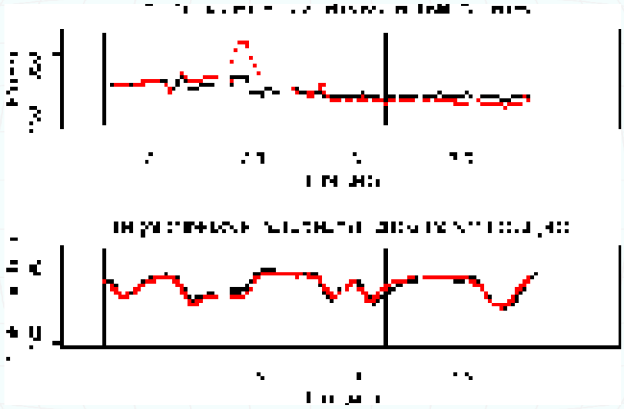
Some of “speech” is visual: shrugs, nods,looking

Question and statement in a sentence with high level tones



F0 tracks are different for ques and statement.
Jaw movements are similar—no diff in emphasis

Normal sentence and the sentence with one digit in narrow focus



Jaw movement shows the effect of emphasis. The magnitude is weak on a glide+mid vowel