CONTROLLING FOR CONFOUNDS IN ONLINE MEASURES OF SENTENCE COMPLEXITY

Marten van Schijndel [1]

July 28, 2015

[1]Department of Linguistics, The Ohio State University

PREACHING TO THE CHOIR

Occurrence frequencies have major influence on sentence processing

Occurrence frequencies have major influence on sentence processing

H_0 demands that we then control for these factors in our studies

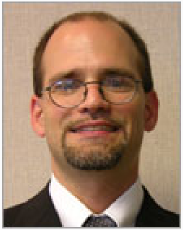
Occurrence frequencies have major influence on sentence processing

H_0 demands that we then control for these factors in our studies

How do people try to account for frequencies?

Case Study 1: Cloze Probabilities
van Schijndel, Culicover, & Schuler (2014)





Pertains to: Pickering & Traxler (2003), inter alia

Ask subjects to generate distribution

Ask subjects to generate distribution

Sentence generation norming:
Write sentences with these words

landed, sneezed, laughed, ...

Ask subjects to generate distribution

Sentence generation norming:
Write sentences with these words

landed, sneezed, laughed, ...

Cloze norming:
Complete this sentence

The pilot landed ________

Ask subjects to generate distribution

Sentence generation norming:
Write sentences with these words

landed, sneezed, laughed, ...

Cloze norming:
Complete this sentence

The pilot landed the plane.

Ask subjects to generate distribution

Sentence generation norming:
Write sentences with these words

landed, sneezed, laughed, ...

Cloze norming:
Complete this sentence

The pilot landed <u>the plane.</u>   The pilot landed <u>in the field.</u>

Ask subjects to generate distribution

Sentence generation norming:
Write sentences with these words

landed, sneezed, laughed, ...

Cloze norming:
Complete this sentence

NP:The pilot landed the plane. PP: The pilot landed in the field.

25% 40%

Pickering & Traxler (2003) used 6 cloze tasks to determine frequencies

pickering & traxler (2003)

Stimuli

(1) That's the plane that the pilot landed behind in the fog.
(2) That's the truck that the pilot landed behind in the fog.

Readers slow down at *landed* in (2)

Pickering & Traxler (2003)

Stimuli

(1) That's the plane that the pilot landed behind in the fog.
(2) That's the truck that the pilot landed behind in the fog.

Readers slow down at *landed* in (2)

Suggests they try to link *truck* as the object of *landed* despite:

- *landed* biased for PP complement
    - 40% PP complement
    - 25% NP complement

Readers initially adopt a transitive interpretation despite subcat bias

pickering & traxler (2003)

Readers initially adopt a transitive interpretation despite subcat bias

$\therefore$ Early-attachment processing heuristic

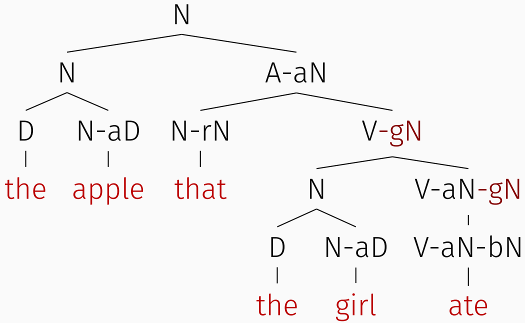
Readers initially adopt a transitive interpretation despite subcat bias

$\therefore$ Early-attachment processing heuristic

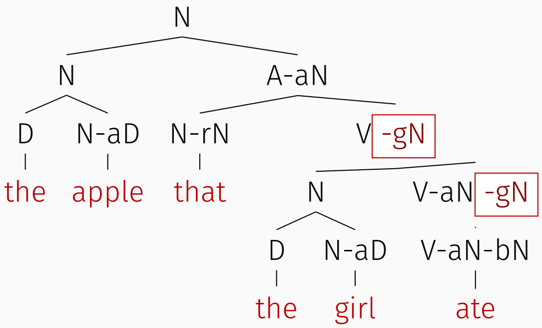
But what about syntactic frequencies?

GENERALIZED CATEGORIAL GRAMMAR (GCG)

Nguyen et al. (2012)



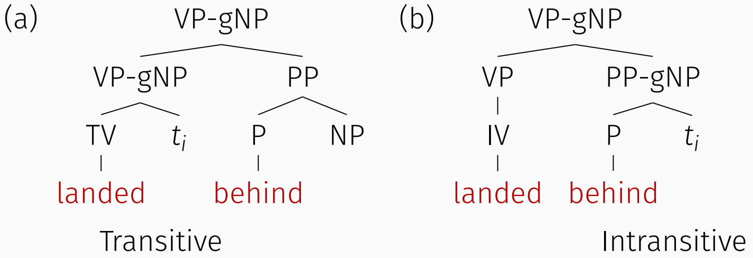
Nguyen et al. (2012)



Pickering & Traxler (2003)

(1) That's the plane that the pilot landed behind in the fog.

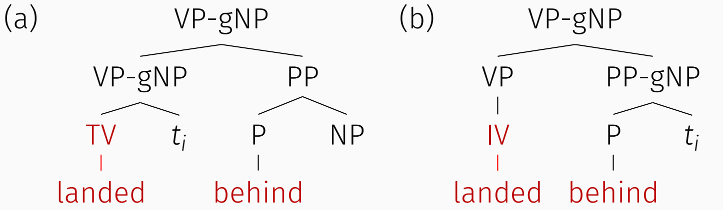
(2) That's the truck that the pilot landed behind in the fog.



Pickering & Traxler (2003)

(1) That's the plane that the pilot landed behind in the fog.

(2) That's the truck that the pilot landed behind in the fog.



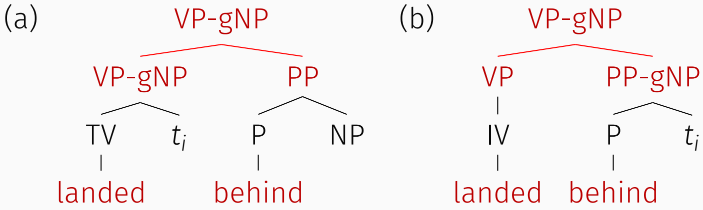
Transitive

Intransitive

Pickering & Traxler (2003)

(1) That's the plane that the pilot landed behind in the fog.

(2) That's the truck that the pilot landed behind in the fog.



Transitive

Intransitive

Pickering & Traxler (2003)

(1) That's the plane that the pilot landed behind in the fog.
(2) That's the truck that the pilot landed behind in the fog.

van Schijndel et al. (2014)
Using syntactic probabilities with cloze data:

Pickering & Traxler (2003)

(1) That's the plane that the pilot landed behind in the fog.
(2) That's the truck that the pilot landed behind in the fog.

van Schijndel et al. (2014)
Using syntactic probabilities with cloze data:

$$P(\text{Transitive} \mid \text{landed}) \propto 0.016$$

$$P(\text{Intransitive} \mid \text{landed}) \propto 0.004$$

WHAT ABOUT SYNTACTIC FREQUENCIES?

Pickering & Traxler (2003)

(1) That's the plane that the pilot landed behind in the fog.
(2) That's the truck that the pilot landed behind in the fog.

van Schijndel et al. (2014)
Using syntactic probabilities with cloze data:

$$P(\text{Transitive} \mid \text{landed}) \propto 0.016$$
$$P(\text{Intransitive} \mid \text{landed}) \propto 0.004$$

Transitive interpretation is 300% more likely!

Subcat processing accounted for by hierarchic syntactic frequencies
Early attachment heuristic unnecessary

Subcat processing accounted for by hierarchic syntactic frequencies
Early attachment heuristic unnecessary

Also applies to heavy-NP shift heuristics (Staub, 2006), unaccusative processing (Staub et al., 2007), etc.

Subcat processing accounted for by hierarchic syntactic frequencies
Early attachment heuristic unnecessary

Also applies to heavy-NP shift heuristics (Staub, 2006), unaccusative processing (Staub et al., 2007), etc.

Suggests cloze probabilities are insufficient as a frequency control

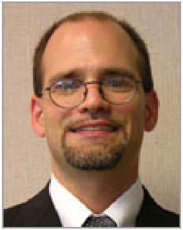
Subcat processing accounted for by hierarchic syntactic frequencies
Early attachment heuristic unnecessary

Also applies to heavy-NP shift heuristics (Staub, 2006), unaccusative processing (Staub et al., 2007), etc.

Suggests cloze probabilities are insufficient as a frequency control

But do people use hierarchic syntactic probabilities?

Case Study 2: *N*-grams and Syntactic Probabilities
van Schijndel & Schuler (2015)

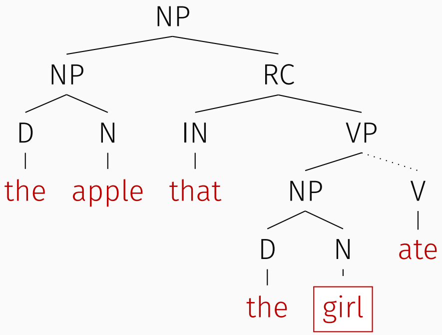


Pertains to: Frank & Bod (2011), inter alia

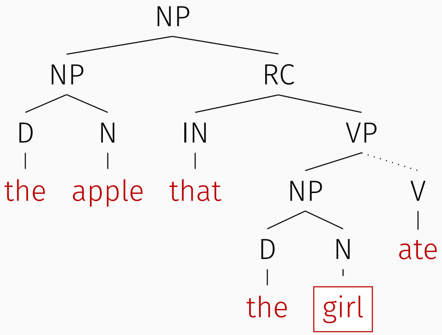
Previous studies have debated whether humans use hierarchic syntax

CASE STUDY 2: OVERVIEW

Previous studies have debated whether humans use hierarchic syntax



Previous studies have debated whether humans use hierarchic syntax



But how robust were their models?

This work shows that:

This work shows that:
N-gram models can be greatly improved (accumulation)

This work shows that:
N-gram models can be greatly improved (accumulation)

Hierarchic syntax is still predictive over stronger baseline

This work shows that:
N-gram models can be greatly improved (accumulation)

Hierarchic syntax is still predictive over stronger baseline

Hierarchic syntax not improved by accumulation

1 2

The red apple that the $\boxed{\text{girl}}$ ate ...

Frank & Bod (2011)

1 2

The red apple that the $\boxed{\text{girl}}$ ate ...

Frank & Bod (2011)

Baseline:

- Sentence Position
- Word length
- *N*-grams (Unigram, bigram)

The red apple that the $\boxed{\text{girl}}$ ate ...
w_1 w_2 w_3 w_4 w_5 w_6

Frank & Bod (2011)

Baseline:

- Sentence Position
- Word length
- *N*-grams (Unigram, bigram)

The red apple that the $\overbrace{\boxed{\text{girl}}}^{\text{4 chars}}_{w_6}$ ate ...

Frank & Bod (2011)

Baseline:

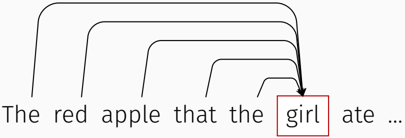
- Sentence Position
- Word length
- *N*-grams (Unigram, bigram)

$$\text{The red apple that } \underline{\text{the}} \; \underbrace{\overbrace{\boxed{\text{girl}}}^{\text{4 chars}}}_{w_6} \; \text{ate} \ldots$$

Frank & Bod (2011)

Baseline:

- Sentence Position
- Word length
- *N*-grams (Unigram, bigram)



Frank & Bod (2011)

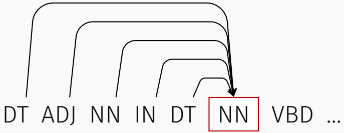
Baseline:

- Sentence Position
- Word length
- *N*-grams (Unigram, bigram)

Test POS Predictors:

- Echo State Network (ESN)
- Phrase Structure Grammar (PSG)

HIERARCHIC SYNTAX IN READING?



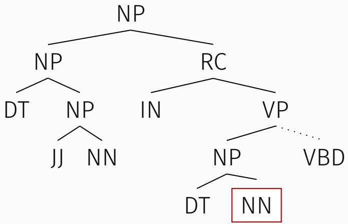
Frank & Bod (2011)

Baseline:

- Sentence Position
- Word length
- *N*-grams (Unigram, bigram)

Test POS Predictors:

- Echo State Network (ESN)
- Phrase Structure Grammar (PSG)



Frank & Bod (2011)

Baseline:

- Sentence Position
- Word length
- *N*-grams (Unigram, bigram)

Test POS Predictors:

- Echo State Network (ESN)
- Phrase Structure Grammar (PSG)

Frank & Bod (2011)

Baseline:

- Sentence Position
- Word length
- *N*-grams (Unigram, bigram)

Test POS Predictors:

- Echo State Network (ESN)
- Phrase Structure Grammar (PSG)

Frank & Bod (2011)

Baseline:

- Sentence Position
- Word length
- *N*-grams (Unigram, bigram)

Test POS Predictors:

- Echo State Network (ESN)
- Phrase Structure Grammar (PSG)

Outcome:

PSG $<$ ESN + PSG

ESN $=$ ESN + PSG

Frank & Bod (2011)

Baseline:

- Sentence Position
- Word length
- *N*-grams (Unigram, bigram)

Test POS Predictors:

- Echo State Network (ESN)
- Phrase Structure Grammar (PSG)

Outcome:

$\text{PSG} < \text{ESN + PSG}$ Sequential helps over hierarchic

$\text{ESN} = \text{ESN + PSG}$

Frank & Bod (2011)

Baseline:

- Sentence Position
- Word length
- *N*-grams (Unigram, bigram)

Test POS Predictors:

- Echo State Network (ESN)
- Phrase Structure Grammar (PSG)

Outcome:

PSG $<$ ESN + PSG

ESN $=$ ESN + PSG   Hierarchic doesn't help over sequential

Fossum & Levy (2012)

Replicated Frank & Bod (2011):

PSG $<$ ESN + PSG

ESN $=$ ESN + PSG

Fossum & Levy (2012)

Replicated Frank & Bod (2011):

PSG $<$ ESN + PSG

ESN $=$ ESN + PSG

Better *n*-gram baseline (more data) changes result:

PSG $\boxed{=}$ ESN + PSG

ESN $=$ ESN + PSG

Fossum & Levy (2012)

Replicated Frank & Bod (2011):

PSG $<$ ESN + PSG

ESN $=$ ESN + PSG

Better *n*-gram baseline (more data) changes result:

PSG $\boxed{=}$ ESN + PSG   Sequential doesn't help over hierarchic

ESN $=$ ESN + PSG

Fossum & Levy (2012)

Replicated Frank & Bod (2011):

PSG $<$ ESN + PSG

ESN $=$ ESN + PSG

Better *n*-gram baseline (more data) changes result:

PSG $\boxed{=}$ ESN + PSG   Sequential doesn't help over hierarchic

ESN $=$ ESN + PSG

Also: lexicalized syntax improves PSG fit

Previous reading time studies:

- Unigrams/Bigrams/Trigrams
  Trained on WSJ, Dundee, BNC

Previous reading time studies:

- Unigrams/Bigrams/Trigrams
  Trained on WSJ, Dundee, BNC
- Only from region boundaries

Bigram Example

Reading time of *girl* after *red*

The $\overset{1}{\text{red}}$ $\underbrace{\text{apple that } \underline{\text{the}} \boxed{\overset{2}{\text{girl}}}}_{\text{region}}$ ate ...

$\boxed{\text{X}}$: bigram target $\underline{\text{X}}$: bigram condition

Improved n-gram baseline

Bigram Example

Reading time of *girl* after *red*

1 2

The red apple that the girl ate ...

region

X: bigram target X: bigram condition

- Fails to capture entire sequence;

Bigram Example

Reading time of *girl* after *red*

The red^1 $\underbrace{\text{apple that } \underline{\text{the}} \boxed{\text{girl}}^2}_{\text{region}}$ ate …

$\boxed{\text{X}}$: bigram target $\underline{\text{X}}$: bigram condition

- Fails to capture entire sequence;
- Conditions never generated;

Bigram Example

Reading time of *girl* after *red*

The red[1] $\underbrace{\text{apple that }\underline{\text{the}}\ \boxed{\text{girl}}^{2}}_{\text{region}}$ ate …

$\boxed{X}$: bigram target $\underline{X}$: bigram condition

- Fails to capture entire sequence;
- Conditions never generated;
- Probability of sequence is deficient

Cumulative Bigram Example

Reading time of *girl* after *red*:

The $\overset{1}{\underline{\text{red}}}$ $\boxed{\underline{\text{apple}}}$ $\boxed{\underline{\text{that}}}$ $\boxed{\underline{\text{the}}}$ $\overset{2}{\boxed{\text{girl}}}$ ate ...

$\boxed{\text{X}}$: bigram targets $\underline{\text{X}}$: bigram conditions

Cumulative Bigram Example

Reading time of *girl* after *red*:

The red ^{1} $\boxed{\text{apple}}$ $\boxed{\text{that}}$ $\boxed{\text{the}}$ $\boxed{\text{girl}}^{2}$ ate ...

$\boxed{\text{X}}$: bigram targets   X: bigram conditions

- Captures entire sequence;
- Well-formed sequence probability;
- Reflects processing that must be done by humans

Previous reading time studies:

- Unigrams/Bigrams/Trigrams
- Trained on WSJ, Dundee, BNC
- Only from region boundaries

Previous reading time studies:

- Unigrams/Bigrams/Trigrams
- Trained on WSJ, Dundee, BNC
- Only from region boundaries

This study:

- 5-grams (w/ backoff)
- Trained on Gigaword 4.0
- Cumulative and Non-cumulative

EVALUATION

Dundee Corpus (Kennedy et al., 2003)

- 10 subjects
- 2,388 sentences
- 58,439 words
- 194,882 first pass durations
- 193,709 go-past durations

Exclusions:

- Unknown words (5 tokens)
- First and last of a line
- Regions larger than 4 words (track loss)

Baseline:

Fixed Effects

- Sentence Position
- Word length
- Region Length
- Preceding word fixated?

Random Effects

- Item/Subject Intercepts
- By Subject Slopes:
  - All Fixed Effects
  - *N*-grams (5-grams)
  - *N*-grams (Cumu-5-grams)

Baseline:

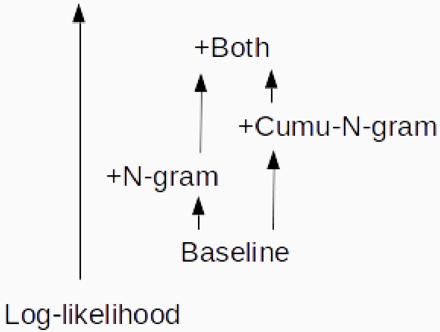
Fixed Effects

- Sentence Position
- Word length
- Region Length
- Preceding word fixated?

Random Effects

- Item/Subject Intercepts
- By Subject Slopes:
  - All Fixed Effects
  - N-grams (5-grams) $\leftarrow$
  - N-grams (Cumu-5-grams) $\leftarrow$

CUMU-n-GRAMS PREDICT READING TIMES



First Pass and Go-Past

- Is hierarchic surprisal useful over the better baseline?

FOLLOW-UP QUESTIONS

- Is hierarchic surprisal useful over the better baseline?
- If so, can it be similarly improved through accumulation?

- Is hierarchic surprisal useful over the better baseline?
- If so, can it be similarly improved through accumulation?
  van Schijndel & Schuler (2013) found it could over weaker baselines

Grammar:
Berkeley parser, WSJ, 5 split-merge cycles (Petrov & Klein 2007)

Baseline:

Fixed Effects

- Same as before
- *N*-grams (5-grams)
- *N*-grams (Cumu-5-grams)

Baseline:

Fixed Effects

- Same as before
- *N*-grams (5-grams)
- *N*-grams (Cumu-5-grams)

Random Effects

- Same as before
- By Subject Slopes:
  - Hierarchic surprisal
  - Cumu-Hierarchic surprisal

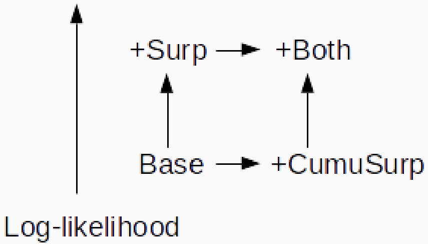
Baseline:

Fixed Effects

- Same as before
- *N*-grams (5-grams)
- *N*-grams (Cumu-5-grams)

Random Effects

- Same as before
- By Subject Slopes:
  - Hierarchic surprisal $\leftarrow$
  - Cumu-Hierarchic surprisal $\leftarrow$



First Pass and Go-Past

- Suggests previous findings were due to weaker *n*-gram baseline

CUMULATIVE SURPRISAL DOESN'T HELP?!

- Suggests previous findings were due to weaker *n*-gram baseline
- Suggests only local PCFG surprisal affects reading times

CUMULATIVE SURPRISAL DOESN'T HELP?!

- Suggests previous findings were due to weaker *n*-gram baseline
- Suggests only local PCFG surprisal affects reading times

Follow-up work shows long distance dependencies independently influence reading times

Hierarchic syntax predicts reading times over strong linear baseline

Hierarchic syntax predicts reading times over strong linear baseline

Studies should use cumu-*n*-grams in their baselines

WHAT DOES THIS MEAN FOR OUR MODELS?

We need to carefully control for:

- Cloze probabilities

WHAT DOES THIS MEAN FOR OUR MODELS?

We need to carefully control for:

- Cloze probabilities
- *N*-gram frequencies (local and cumulative)

We need to carefully control for:

- Cloze probabilities
- *N*-gram frequencies (local and cumulative)
- Hierarchic syntactic frequencies

We need to carefully control for:

- Cloze probabilities
- *N*-gram frequencies (local and cumulative)
- Hierarchic syntactic frequencies
- Long distance dependency frequencies

WHAT DOES THIS MEAN FOR OUR MODELS?

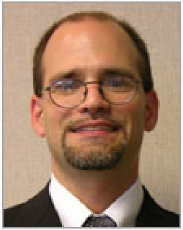
We need to carefully control for:

- Cloze probabilities
- *N*-gram frequencies (local and cumulative)
- Hierarchic syntactic frequencies
- Long distance dependency frequencies
- …(discourse, etc.)

Then we can try to interpret experimental results.

What do we do about convergence?
Is there a way to avoid this explosion of control predictors?

Case Study 3: Evading Frequency Confounds
van Schijndel, Murphy, & Schuler (2015)



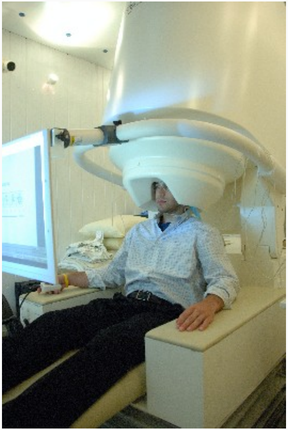
MOTIVATING QUESTION

Can we measure memory load without controlling for frequency effects?

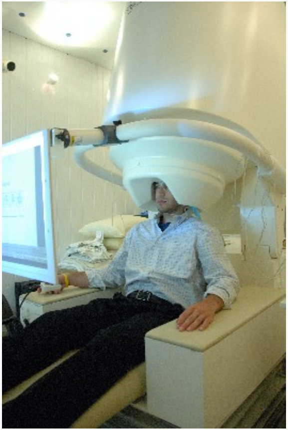
Can we measure memory load without controlling for frequency effects?

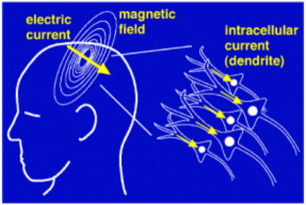
Let's try using MEG.

WHAT IS MEG?



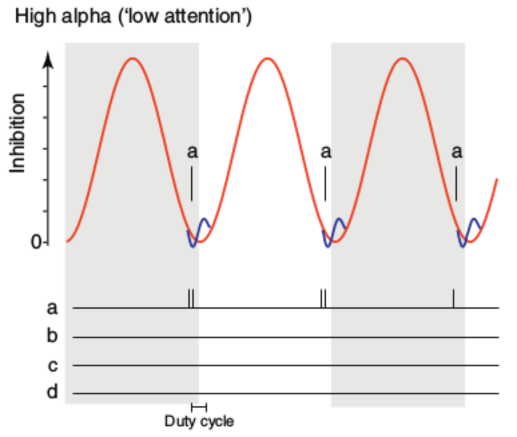
WHAT IS MEG?



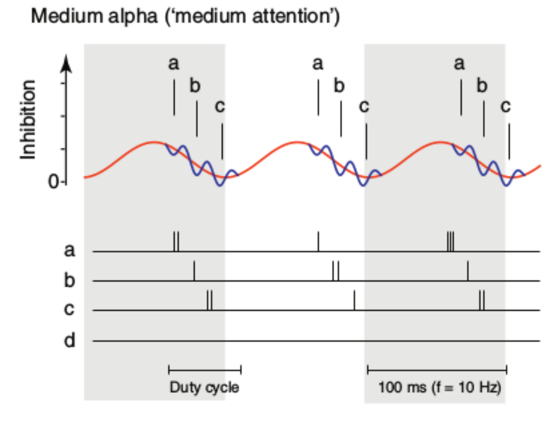


102 locations

HOW MIGHT MEG REFLECT LOAD?



Jensen et al., (2012)



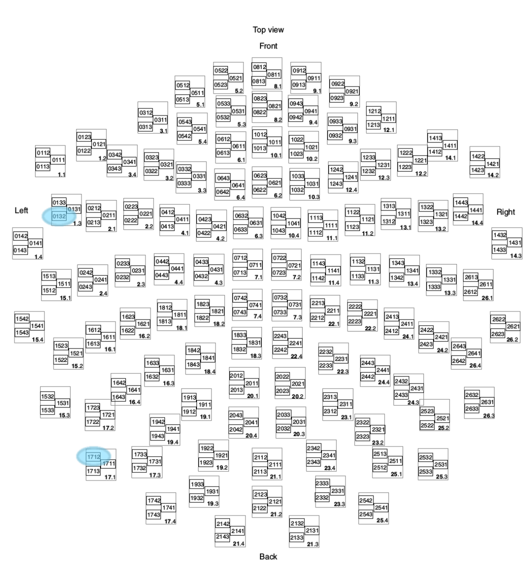
Jensen et al., (2012)

WHERE TO LOOK?

Memory is a function of distributed processing

Memory is a function of distributed processing

Look for synchronized firing between sensors (brain regions)


Top view
Front
Left
Right
Back

Memory is a function of distributed processing

Look for synchronized firing between sensors (brain regions)

This study uses *spectral coherence* measurements.

SPECTRAL COHERENCE

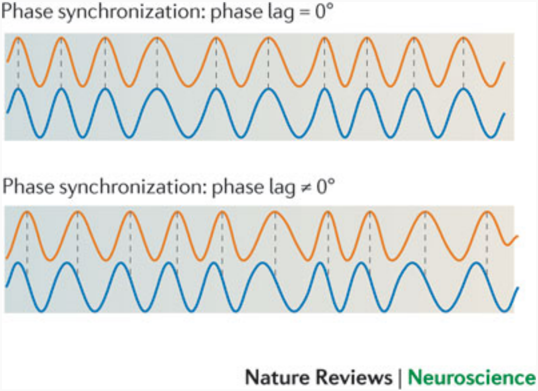
$$\text{coherence}(x, y) = \frac{\mathrm{E}[S_{xy}]}{\sqrt{\mathrm{E}[S_{xx}] \cdot \mathrm{E}[S_{yy}]}}$$

SPECTRAL COHERENCE

$$\text{coherence}(x, y) = \frac{\mathrm{E}[S_{xy}]}{\sqrt{\mathrm{E}[S_{xx}] \cdot \mathrm{E}[S_{yy}]}} \quad \begin{array}{l} \leftarrow \text{cross-correlation} \\ \leftarrow \text{autocorrelations} \end{array}$$

$$\text{coherence}(x, y) = \frac{E[S_{xy}]}{\sqrt{E[S_{xx}] \cdot E[S_{yy}]}} \quad \begin{array}{l} \leftarrow \text{cross-correlation} \\ \leftarrow \text{autocorrelations} \end{array}$$

Amount of connectivity (synchronization) not caused by chance



Fell & Axmacher (2011)

Collected 2 years ago at CMU

AUDIOBOOK MEG CORPUS

Collected 2 years ago at CMU

3 subjects

AUDIOBOOK MEG CORPUS

Collected 2 years ago at CMU

3 subjects

Heart of Darkness, ch. 2

12,342 words

80 (8 x 10) minutes

Synched with parallel audio recording
and forced alignment

AUDIOBOOK MEG CORPUS

Collected 2 years ago at CMU

3 subjects

Heart of Darkness, ch. 2
12,342 words
80 (8 x 10) minutes
Synched with parallel audio recording
and forced alignment

306-channel Elekta Neuromag, CMU
Movement/noise correction: SSP, SSS, tSSS
Band-pass filtered 0.01–50 Hz
Downsampled to 125 Hz
Visually scanned for muscle artifacts; none found

*d*1 The cart broke.
*d*2 that the man bought

*d*1	The cart		broke.
*d*2		that the man bought	

Depth annotations:
van Schijndel et al., (2013) parser
Nguyen et al., (2012) Generalized Categorial Grammar (GCG)

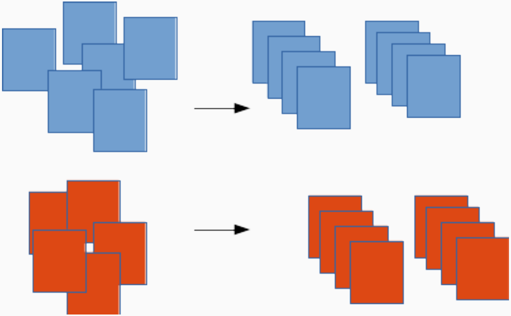
Remove words:

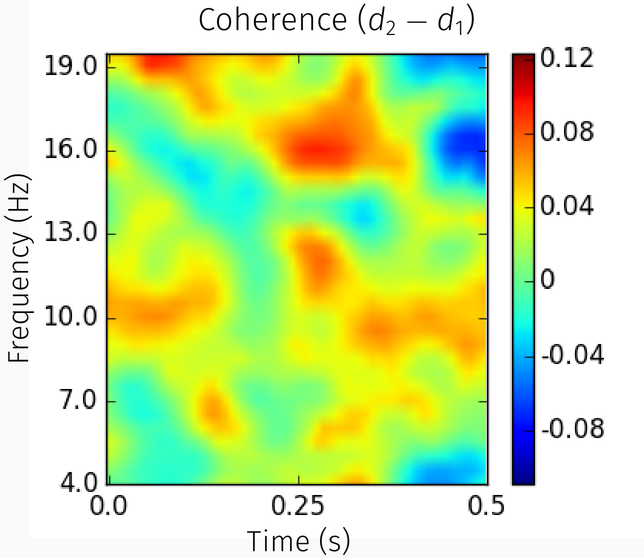
- in short or long sentences ($<$4 or $>$50 words)
- that follow a word at another depth
- that fail to parse

Partition data:

- Dev set: One third of corpus
- Test set: Two thirds of corpus

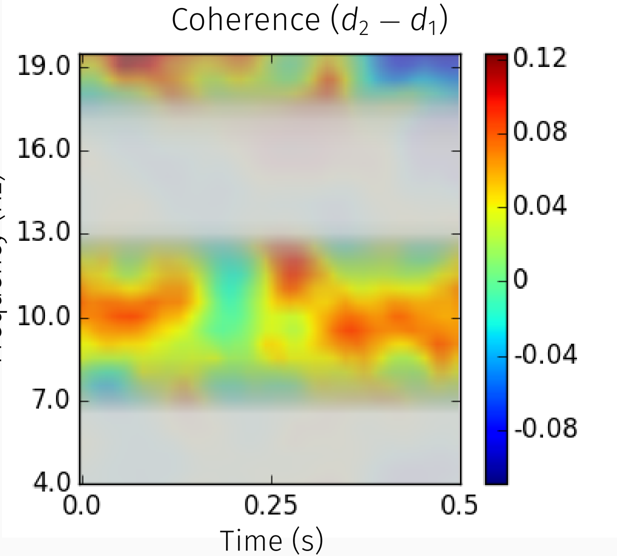
- Group by factor
- Compute coherence over subsets of 4 epochs





Coherence ($d_2 - d_1$)
19.0
16.0
13.0
10.0
7.0
4.0
Frequency (Hz)
0.0
0.25
0.5
Time (s)
0.12
0.08
0.04
0
-0.04
-0.08

DEV COHERENCE +VARIANCE



POSSIBLE CONFOUNDS?

Sentence position

Unigram, Bigram, Trigram: COCA logprobs

PCFG surprisal: parser output

Factor	p-value
Unigram	0.941
Bigram	0.257
Trigram	0.073
PCFG Surprisal	0.482
Sentence Position	0.031
Depth	0.005

Depth 1 (40 items) Depth 2 (1118 items)

TEST RESULTS

Factor	p-value
Unigram	0.6480
Bigram	0.7762
Trigram	0.0264
PCFG Surprisal	0.3295
Sentence Position	0.4628
Depth	0.00002

Depth 1 (86 items)   Depth 2 (2142 items)

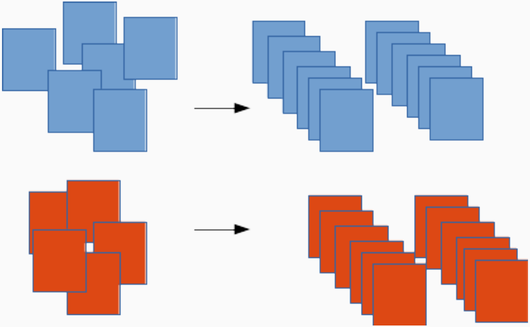
TEST RESULTS

Factor	p-value
Unigram	0.6480
Bigram	0.7762
Trigram	0.0264
PCFG Surprisal	0.3295
Sentence Position	0.4628
Depth	0.00002

Bonferroni correction removes trigrams, but ...

COMPUTE COHERENCE: INCREASED RESOLUTION

- Group by factor
- Compute coherence over subsets of 6 epochs



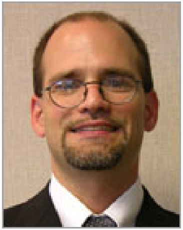
Factor	p-value
Trigram	0.3817
Depth	0.0046

Depth 1 (57 items)   Depth 2 (1428 items)

- Memory load is reflected in MEG connectivity
- Common confounds do not pose problems for oscillatory measures

- Cloze probabilities are insufficient as frequency control
- Hierarchic syntactic frequencies strongly influence processing
- Reading time studies need to use local *and* cumulative *n*-grams
- Oscillatory analyses could avoid control predictor explosion





- Stefan Frank, Matthew Traxler, Shari Speer, Roberto Zamparelli
- Attendees of CogSci 2014, CUNY 2015, NAACL 2015, CMCL 2015
- OSU Linguistics Targeted Investment for Excellence (2012-2013)
- National Science Foundation (DGE-1343012)
- University of Pittsburgh Medical Center MEG Seed Fund
- National Institutes of Health CRCNS (5R01HD075328-02)

- Cloze probabilities are insufficient as frequency control
- Hierarchic syntactic frequencies strongly influence processing
- Reading time studies need to use local *and* cumulative *n*-grams
- Oscillatory analyses could avoid control predictor explosion