FREQUENCIES THAT MATTER IN SENTENCE PROCESSING

Marten van Schijndel [1]

September 7, 2015

[1]Department of Linguistics, The Ohio State University

FREQUENCIES ARE IMPORTANT

Occurrence frequencies describe languages well

- Zipf
- Statistical NLP (esp. vector spaces)

Occurrence frequencies describe languages well

- Zipf
- Statistical NLP (esp. vector spaces)

Occurrence frequencies have major influence on sentence processing

- Behavioral measures (e.g., reading times)
- Processing measures (e.g., ERPs)
- Uniform Information Density
- Saarland SFB-1102

NULL HYPOTHESIS DEMANDS CONTROL

NULL HYPOTHESIS DEMANDS CONTROL

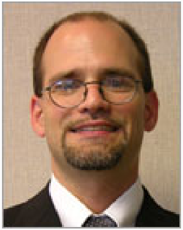
Linguists must control for frequencies to do research

NULL HYPOTHESIS DEMANDS CONTROL

Linguists must control for frequencies to do research

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

GENERATION/CLOZE PROBABILITIES

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.

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

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

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)

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

pickering & traxler (2003)

Readers initially adopt a transitive interpretation despite subcat bias

Readers initially adopt a transitive interpretation despite subcat bias

∴ Early-attachment processing heuristic

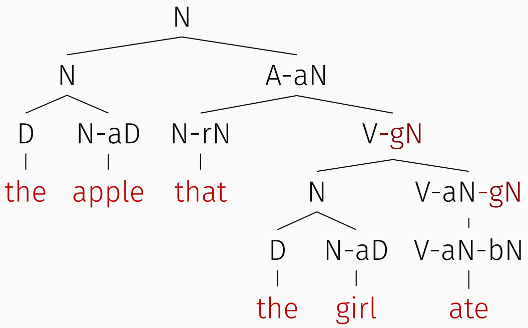
PICKERING & TRAXLER (2003)

Readers initially adopt a transitive interpretation despite subcat bias

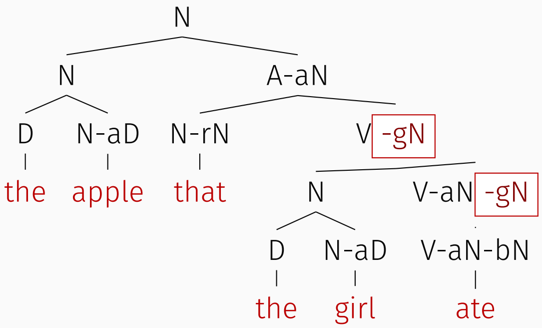
∴ Early-attachment processing heuristic

But what about syntactic frequencies?

Nguyen et al. (2012)



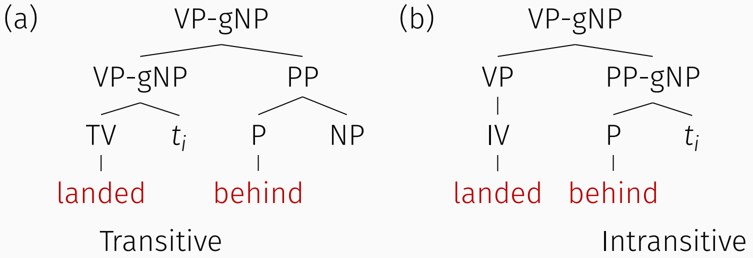
Nguyen et al. (2012)



Pickering & Traxler (2003)

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

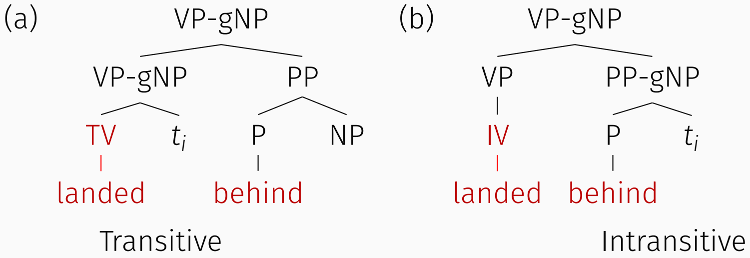
(2) That's the truck that the pilot landed behind in the fog.



Pickering & Traxler (2003)

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

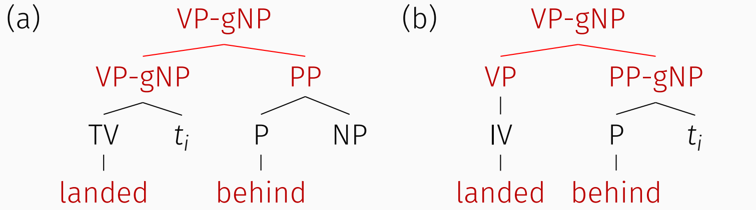
(2) That's the truck that the pilot landed behind in the fog.



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:

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:

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

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

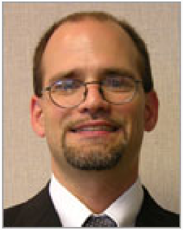
Subcat processing accounted for by hierarchic syntactic frequencies
Early attachment heuristic unnecessary

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?

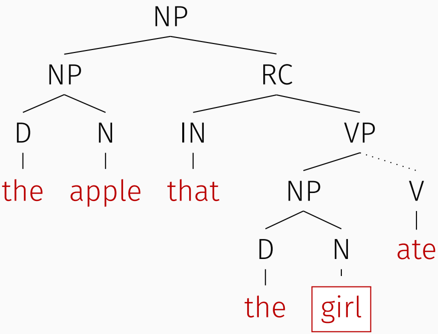
Case Study 2: *N*-grams and Syntactic Probabilities
van Schijndel & Schuler (2015)

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

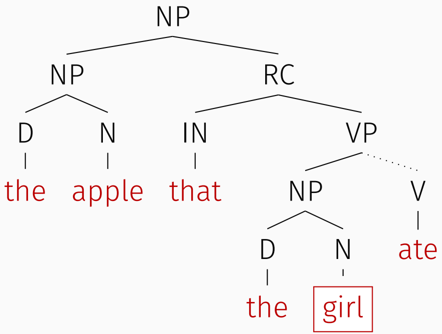
Previous studies have debated whether humans use hierarchic syntax

Previous studies have debated whether humans use hierarchic syntax



CASE STUDY 2: OVERVIEW

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

(Long distance dependencies independently improve model)

1 2

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

FRANK & BOD (2011)

The red[1] apple that the $\boxed{\text{girl}}$[2] 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:

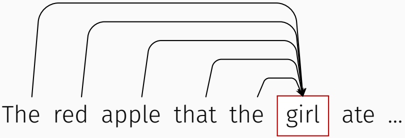
- Sentence Position
- Word length
- *N*-grams (Unigram, bigram)

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

FRANK & BOD (2011)

Baseline:

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



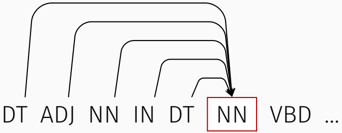
Frank & Bod (2011)

Baseline:

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

Test POS Predictors:

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



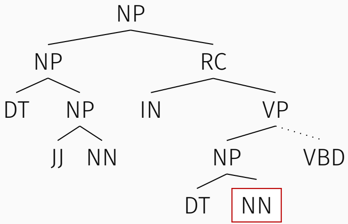
Frank & Bod (2011)

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:

PSG $<$ ESN + PSG Sequential helps over hierarchic

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:

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

Most previous reading time studies:

- *N*-grams trained on WSJ, Dundee, BNC (or less)

IMPROVED n-GRAM BASELINE

Most previous reading time studies:

- *N*-grams trained on WSJ, Dundee, BNC (or less)

This study:

- *N*-grams trained on Gigaword 4.0

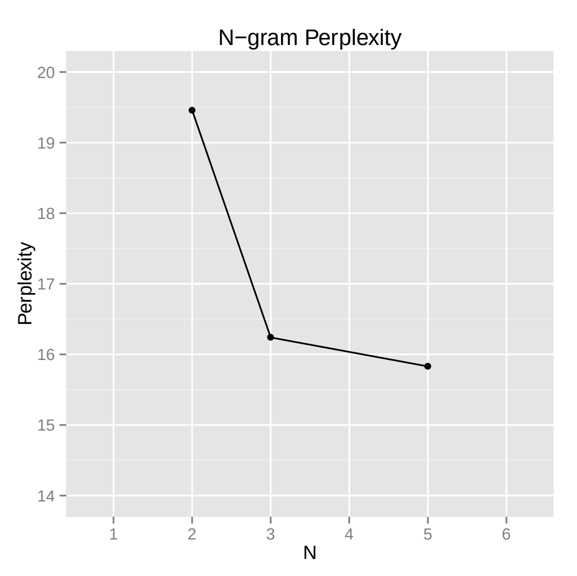
Most previous reading time studies:

- *N*-grams trained on WSJ, Dundee, BNC (or less)
- Unigrams/Bigrams

This study:

- *N*-grams trained on Gigaword 4.0

IMPROVED n-GRAM BASELINE



IMPROVED n-GRAM BASELINE

Most previous reading time studies:

- *N*-grams trained on WSJ, Dundee, BNC
- Unigrams/Bigrams

This study:

- *N*-grams trained on Gigaword 4.0
- 5-grams

IMPROVED n-GRAM BASELINE

Most previous reading time studies:

- *N*-grams trained on WSJ, Dundee, BNC
- Unigrams/Bigrams
- Only from region boundaries

This study:

- *N*-grams trained on Gigaword 4.0
- 5-grams

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

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;

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 …

$\boxed{X}$: bigram target $\underline{X}$: 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 $\underline{\text{red}}^{1}$ $\boxed{\underline{\text{apple}}}$ $\boxed{\underline{\text{that}}}$ $\boxed{\underline{\text{the}}}$ $\boxed{\text{girl}}^{2}$ ate ...

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

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

Most previous reading time studies:

- *N*-grams trained on WSJ, Dundee, BNC
- Unigrams/Bigrams
- Only from region boundaries

This study:

- *N*-grams trained on Gigaword 4.0
- 5-grams

Most previous reading time studies:

- *N*-grams trained on WSJ, Dundee, BNC
- Unigrams/Bigrams
- Only from region boundaries

This study:

- *N*-grams trained on Gigaword 4.0
- 5-grams
- Cumulative and Non-cumulative

Dundee Corpus (Kennedy et al., 2003)

- 10 subjects
- 2,388 sentences
- First pass durations ($\sim$ 200,000)
- Go-past durations ($\sim$ 200,000)

Exclusions:

- Unknown words (<5 tokens)
- First and last of each 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)

CUMU-n-GRAMS PREDICT READING TIMES

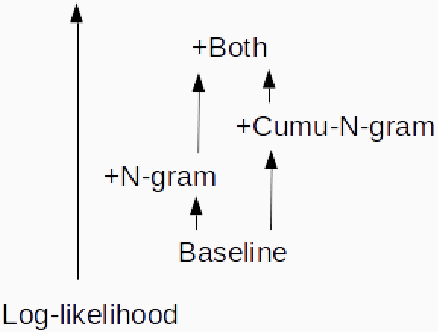
Baseline:

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) $\leftarrow$



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?

FOLLOW-UP QUESTIONS

- 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

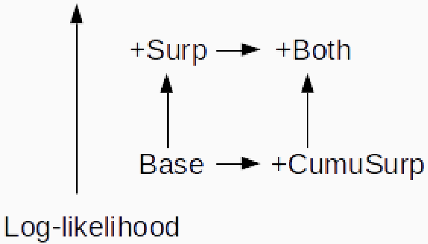
Baseline:

Fixed Effects

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

Random Effects

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



First Pass and Go-Past

CUMULATIVE SURPRISAL DOESN'T HELP?!

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

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

- 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

Long-distance dependencies help over hierarchic syntax

Hierarchic syntax predicts reading times over strong linear baseline

Long-distance dependencies help over hierarchic syntax

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

We need to carefully control for:

- Cloze probabilities (Smith 2011)

We need to carefully control for:

- Cloze probabilities (Smith 2011)
- *N*-gram frequencies (local and cumulative)

WHAT DOES THIS MEAN FOR OUR MODELS?

We need to carefully control for:

- Cloze probabilities (Smith 2011)
- *N*-gram frequencies (local and cumulative)
- Hierarchic syntactic frequencies

WHAT DOES THIS MEAN FOR OUR MODELS?

We need to carefully control for:

- Cloze probabilities (Smith 2011)
- *N*-gram frequencies (local and cumulative)
- Hierarchic syntactic frequencies
- Long distance dependency frequencies

WHAT DOES THIS MEAN FOR OUR MODELS?

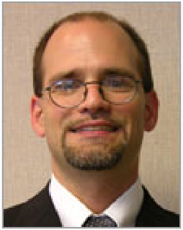
We need to carefully control for:

- Cloze probabilities (Smith 2011)
- *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)



Can we measure memory load with fewer controls?

Can we measure memory load with fewer controls?

Why do so many factors influence results?

Can we measure memory load with fewer controls?

Why do so many factors influence results?

Low dimensionality measures.

Can we measure memory load with fewer controls?

Why do so many factors influence results?

Low dimensionality measures.

Do the factors become separable in another space?

Can we measure memory load with fewer controls?

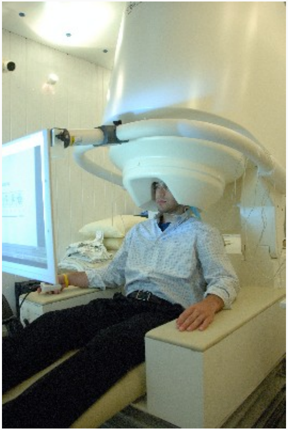
Why do so many factors influence results?

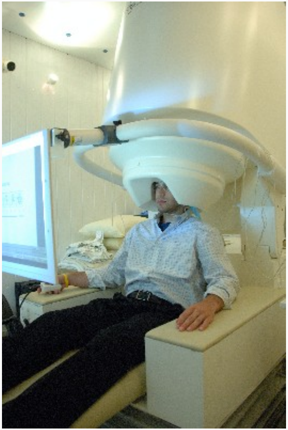
Low dimensionality measures.

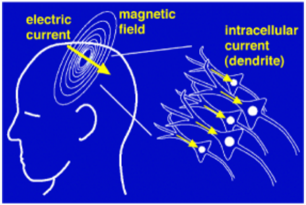
Do the factors become separable in another space?

Let's try using MEG.

WHAT IS MEG?

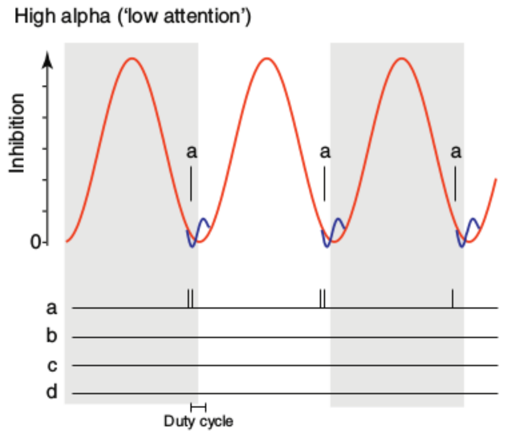






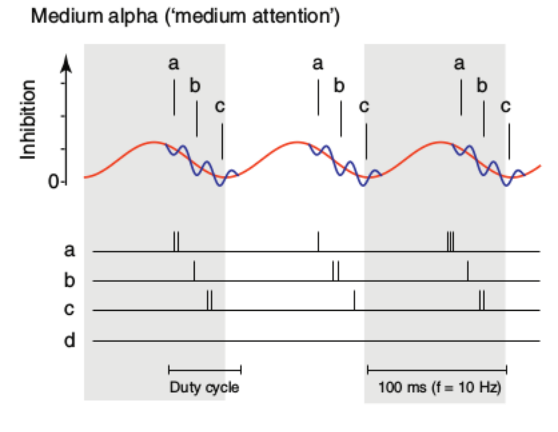
102 locations

HOW MIGHT MEG REFLECT LOAD?



Jensen et al., (2012)

HOW MIGHT MEG REFLECT LOAD?



Jensen et al., (2012)

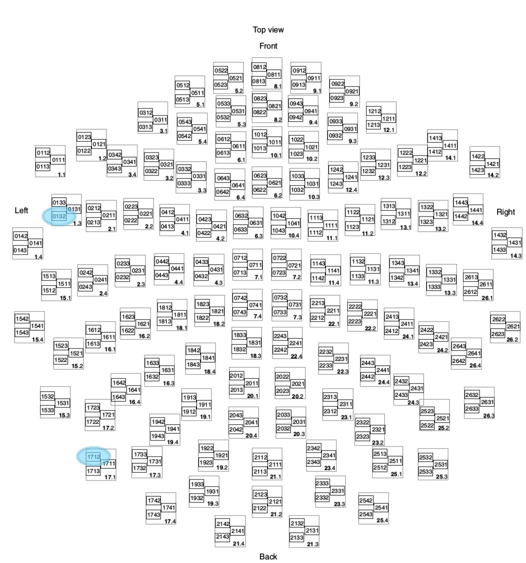
WHERE TO LOOK?

Memory is a function of distributed processing

Memory is a function of distributed processing

Look for synchronized firing between sensors (brain regions)

WHERE TO LOOK?



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

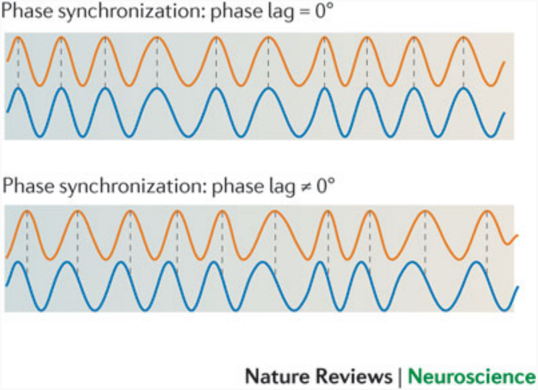
$$\text{coherence}(x, y) = \frac{\mathrm{E}[S_{xy}]}{\sqrt{\mathrm{E}[S_{xx}] \cdot \mathrm{E}[S_{yy}]}} \quad \begin{array}{l} \leftarrow \text{cross-correlation} \\ \leftarrow \text{autocorrelations} \end{array}$$

SPECTRAL COHERENCE

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

Amount of connectivity (synchronization) not caused by chance

SPECTRAL COHERENCE: PHASE SYNCHRONY



Fell & Axmacher (2011)

AUDIOBOOK MEG CORPUS

Collected 2 years ago at CMU

AUDIOBOOK MEG CORPUS

Collected 2 years ago at CMU

3 subjects

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

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)

DATA FILTERING

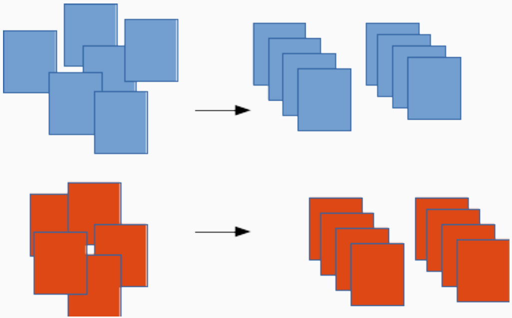
Remove words:

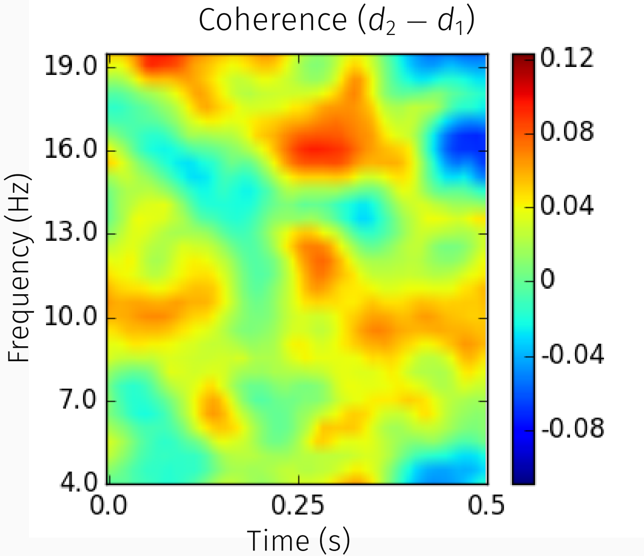
- in short or long sentences (<4 or >50 words)
- that follow a word at another depth
- that fail to parse

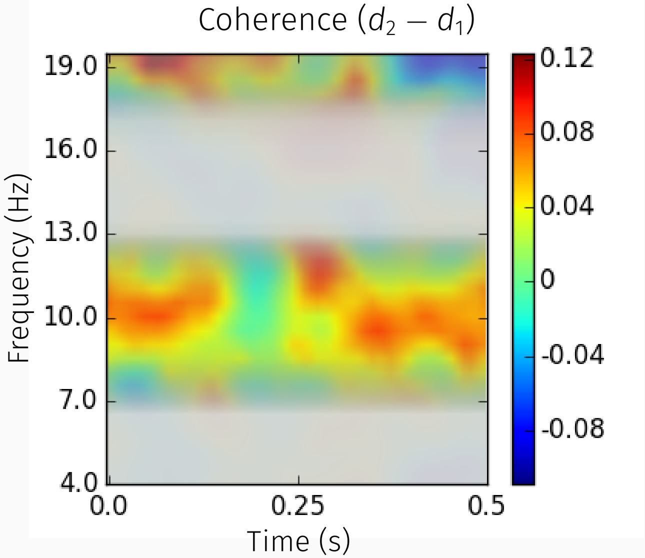
Partition data:

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

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




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


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

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

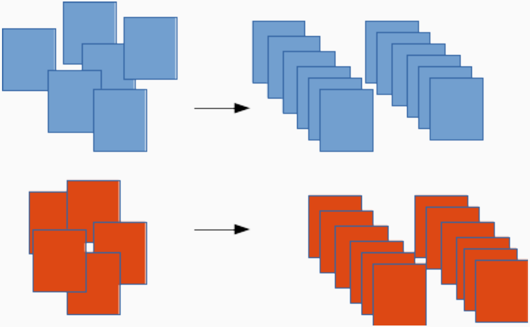
Factor	p-value
Unigram	0.6480
Bigram	0.7762
Trigram	0.0264
PCFG Surprisal	0.3295
Sentence Position	0.4628
Depth	0.00002

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

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 …

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



TEST RESULTS: INCREASED RESOLUTION

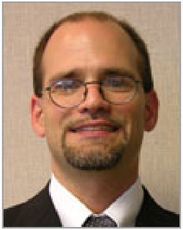
Factor	p-value
Trigram	0.3817
Depth	0.0046

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

ACKNOWLEDGEMENTS

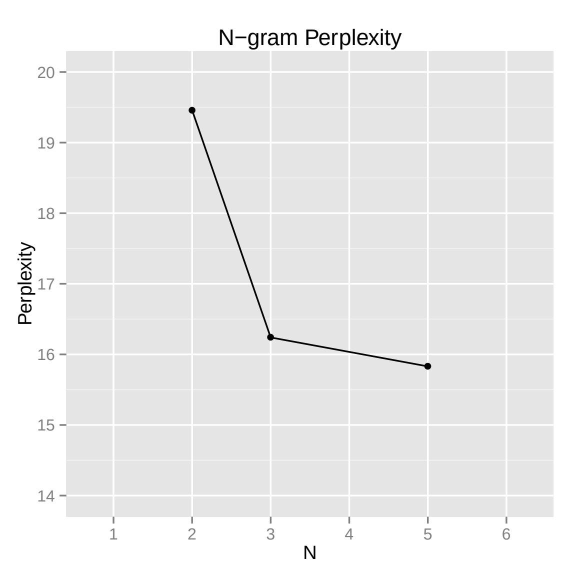
  

- Stefan Frank, Matthew Traxler, Shari Speer, Roberto Zamparelli
- Attendees of CogSci 2014, CUNY 2015, NAACL 2015, CMCL 2015
- OSU Linguistics Targeted Investment for Excellence (2012-2013)
- National Science Foundation (DGE-1343012)
- University of Pittsburgh Medical Center MEG Seed Fund
- National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

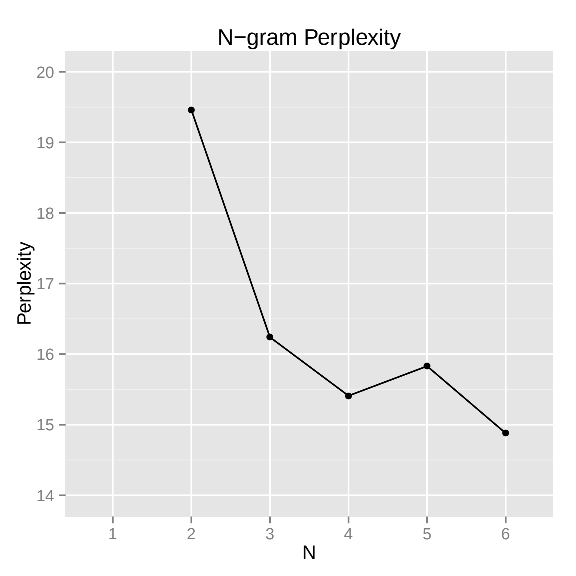
THANKS! QUESTIONS?

- 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

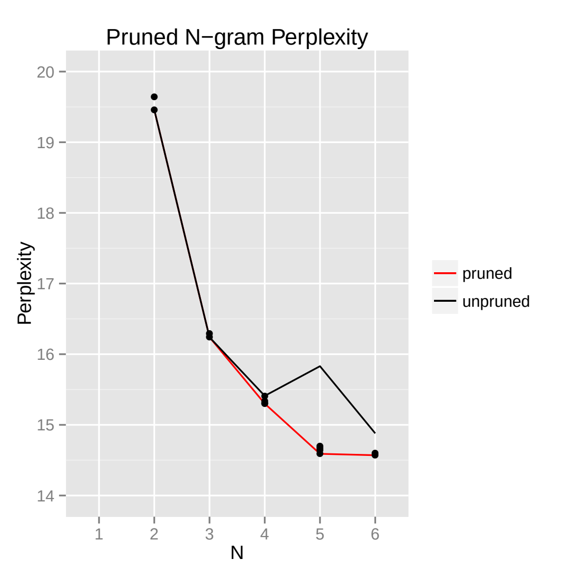
IMPROVED n-GRAM BASELINE



FURTHER IMPROVED n-GRAM BASELINE

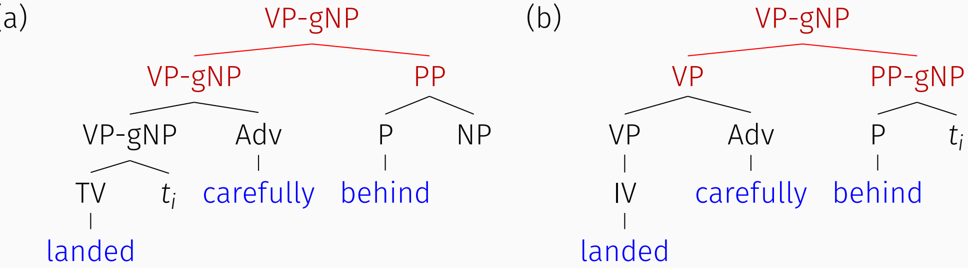


FURTHER IMPROVED n-GRAM BASELINE



How probable is each subtree?

Wall Street Journal (WSJ) section of the Penn Treebank:

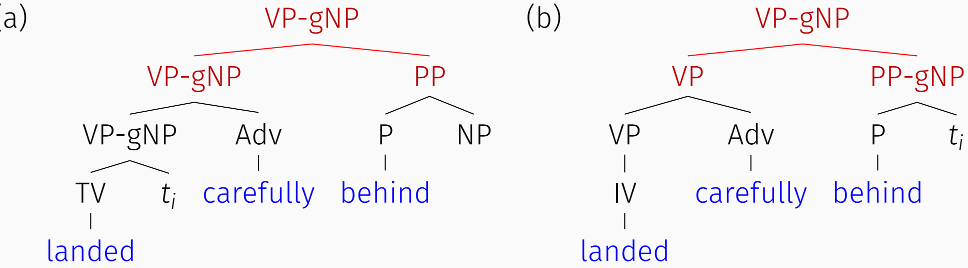


Transitive

Intransitive

How probable is each subtree?

Wall Street Journal (WSJ) section of the Penn Treebank:



Transitive	Intransitive
0.17	0.01

What is the probability of each interpretation?

P(syntactic configuration)·P(generating the verb from that tree)

$$\mathrm{P(Transitive)} = \mathrm{P(VP\text{-}gNP \rightarrow VP\text{-}gNP\ \ PP)} \cdot \mathrm{P}(\mathit{verb} \mid \mathrm{TV}) \tag{1}$$

$$\mathrm{P(Intransitive)} = \mathrm{P(VP\text{-}gNP \rightarrow VP\ \ PP\text{-}gNP)} \cdot \mathrm{P}(\mathit{verb} \mid \mathrm{IV}) \tag{2}$$

What is the probability of each interpretation?

P(syntactic configuration)·~~P(generating the verb from that tree)~~

P(subcat bias)/P(preterminal prior)

$$P(\text{Transitive}) = P(\text{VP-gNP}\rightarrow\text{VP-gNP}\ \ \text{PP}) \cdot P(\textit{verb} \mid \text{TV}) \quad (1)$$

$$P(\text{Intransitive}) = P(\text{VP-gNP}\rightarrow\text{VP}\ \ \text{PP-gNP}) \cdot P(\textit{verb} \mid \text{IV}) \quad (2)$$

What is the probability of each interpretation?

P(syntactic configuration)·P(subcat bias)/P(preterminal prior)

$$\begin{aligned} \mathrm{P}(\mathrm{Transitive}) &= \mathrm{P}(\text{VP-gNP}\rightarrow\text{VP-gNP PP}) \cdot \mathrm{P}(\mathit{verb} \mid \mathrm{TV}) && (1) \\ &\propto \mathrm{P}(\text{VP-gNP}\rightarrow\text{VP-gNP PP}) \cdot \frac{\mathrm{P}(\mathrm{TV} \mid \mathit{verb})}{\mathrm{P}(\mathrm{TV})} \\ \mathrm{P}(\mathrm{Intransitive}) &= \mathrm{P}(\text{VP-gNP}\rightarrow\text{VP PP-gNP}) \cdot \mathrm{P}(\mathit{verb} \mid \mathrm{IV}) && (2) \\ &\propto \mathrm{P}(\text{VP-gNP}\rightarrow\text{VP PP-gNP}) \cdot \frac{\mathrm{P}(\mathrm{IV} \mid \mathit{verb})}{\mathrm{P}(\mathrm{IV})} \end{aligned}$$

What are the preterminal priors?

Relative prior probability from the WSJ:

$$P(\text{TV}) : P(\text{IV}) = 2.6 : 1$$

What is the probability of each interpretation?

P(syntactic configuration)·P(subcat bias)/P(preterminal prior)

$$\text{P(Transitive)} \propto \text{P(VP-gNP}\rightarrow\text{VP-gNP PP)} \cdot \frac{\text{P(TV} \mid \textit{verb}\text{)}}{\text{P(TV)}}$$

$$= 0.17 \cdot \frac{\text{P(TV} \mid \textit{verb}\text{)}}{2.6} \tag{1}$$

$$\text{P(Intransitive)} \propto \text{P(VP-gNP}\rightarrow\text{VP PP-gNP)} \cdot \frac{\text{P(IV} \mid \textit{verb}\text{)}}{\text{P(IV)}}$$

$$= 0.01 \cdot \frac{\text{P(IV} \mid \textit{verb}\text{)}}{1.0} \tag{2}$$

What is the probability of each interpretation?

P(syntactic configuration)·P(subcat bias)/P(preterminal prior)

$$\begin{aligned}
\text{P(Transitive)} &\propto \text{P(VP-gNP}\rightarrow\text{VP-gNP PP)} \cdot \frac{\text{P(TV} \mid \textit{verb}\text{)}}{\text{P(TV)}} \\
&= 0.17 \cdot \frac{\text{P(TV} \mid \textit{verb}\text{)}}{2.6} = 0.065 \cdot \text{P(TV} \mid \textit{verb}\text{)} \qquad (1) \\
\text{P(Intransitive)} &\propto \text{P(VP-gNP}\rightarrow\text{VP PP-gNP)} \cdot \frac{\text{P(IV} \mid \textit{verb}\text{)}}{\text{P(IV)}} \\
&= 0.01 \cdot \frac{\text{P(IV} \mid \textit{verb}\text{)}}{1.0} = 0.01 \cdot \text{P(IV} \mid \textit{verb}\text{)} \qquad (2)
\end{aligned}$$

What is the probability of each interpretation?

P(syntactic configuration)·P(subcat bias)/P(preterminal prior)

$$
\begin{aligned}
\text{P(Transitive)} &\propto \text{P(VP-gNP}\rightarrow\text{VP-gNP PP)} \cdot \frac{\text{P(TV} \mid \textit{verb}\text{)}}{\text{P(TV)}} \\
&= 0.17 \cdot \frac{\text{P(TV} \mid \textit{verb}\text{)}}{2.6} = 0.065 \cdot \text{P(TV} \mid \textit{verb}\text{)} \qquad (1) \\
\text{P(Intransitive)} &\propto \text{P(VP-gNP}\rightarrow\text{VP PP-gNP)} \cdot \frac{\text{P(IV} \mid \textit{verb}\text{)}}{\text{P(IV)}} \\
&= 0.01 \cdot \frac{\text{P(IV} \mid \textit{verb}\text{)}}{1.0} = 0.01 \cdot \text{P(IV} \mid \textit{verb}\text{)} \qquad (2)
\end{aligned}
$$

Pickering & Traxler (2003) experimentally determined subcat biases for a wide variety of verbs

Pickering & Traxler (2003)

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

Using Pickering & Traxler's (2003) subcat bias data:

Pickering & Traxler (2003)

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

Using Pickering & Traxler's (2003) subcat bias data:

$$P(\text{Transitive} \mid \text{landed}) \propto 0.17 \cdot \frac{0.25}{2.6} = 0.016$$
$$P(\text{Intransitive} \mid \text{landed}) \propto 0.01 \cdot \frac{0.40}{1.0} = 0.004$$