Frequency and multi-word sequences: A psycholinguistic comparison of two corpora

Benjamin V. Tucker
University of Alberta, Canada
bvtucker@ualberta.ca

Antoine Tremblay
Georgetown University
trea26@gmail.com

Awknowledgements

Shannon Lemke, Cacilia Gagnon, Janelle Dickout, Sheila Charlton, Ross Munro, Kim Wong, and Michelle Sims

SSHRC/CRSH dissertation grant

Background

- Higher 4-word chunk frequency (*in the middle of*) speeds-up reading and facilitates learning in recall tasks (Tremblay et al., 2009)
- 4-word chunk frequency also affects early brain waves (Tremblay & Baayen, in press)
- This experiment investigates effects of chunk frequency on production of Canadian English
  - Ideally, we would use frequency counts from a corpus of Canadian English (ICE Canada)
  - But at time of experimentation, there was British National Corpus (BNC: Davies, 2004) and soon thereafter Contemporary Corpus of American English (COCA: Davies, 2008).

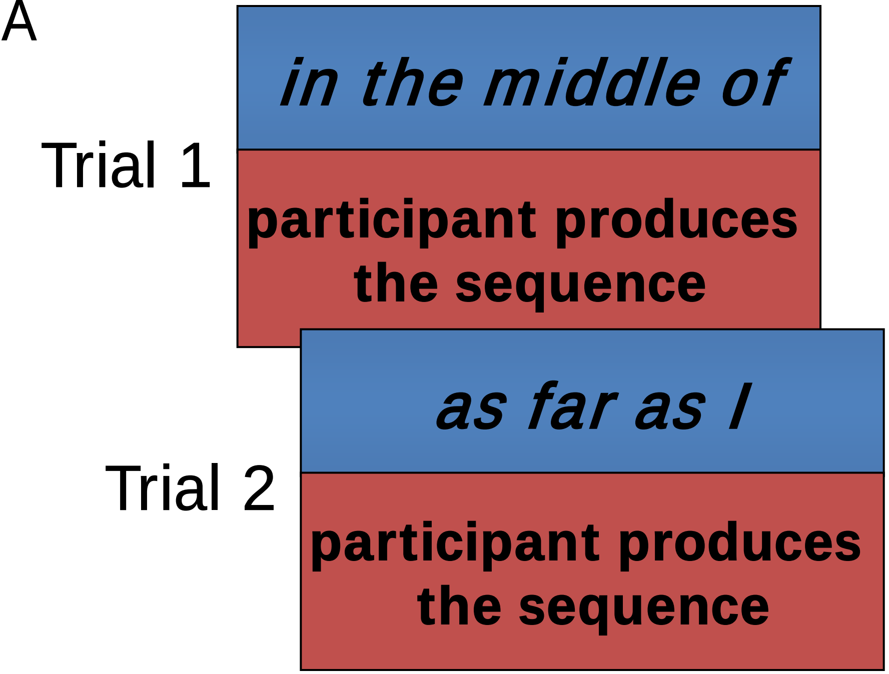
Research Question

- How would the statistical analysis of our data vary as a function of frequencies from the BNC and from COCA?
- Would COCA be more accurate than BNC
  - Canadian English is closer to COCA than BNC
  - Intuitively we would expect that frequency counts from various corpora of English overlap, at least to some extent
  - But there are dialectal differences regarding frequency of use for word and sequences
- We analyzed our data with BNC and COCA counts to see what happens

Experiment

- 432 four-word sequences (chunks) with frequencies were presented to participants one at a time.
  - 0.01 to 100 per million in BNC
  - 0.03 to 85 per million in COCA

- Participants were asked to produce each sequence as soon as they saw them.



Trial 1

in the middle of

participant produces the sequence

Trial 2

as far as I

participant produces the sequence

Measurements

- The time from onset of visual sequence on the screen to the onset of production (response latency) was measured for 24 participants.

- The duration of each sequence (production duration) and any mispronunciations were also measured for 17 of the participants.

Experiment

- Frequencies extracted from:
  - British National Corpus (Davies, 2004)
  - Contemporary Corpus of American English (Davies, 2008).
- Linear mixed-effects regression (LMER) analysis (Baayen, 2008).
- All errors in production (e.g. wrong word) have been removed from the analysis (approx. 20%).

Potential Predictors of Performance

- **Length**

  Number of Letters

  Number of Syllables

- **Token Frequency**

  (from BNC, per million)

  Frequency A, B, C, D

  Frequency AB, BC, CD

  Frequency ABC, BCD

  Frequency ABCD

- **Phrasehood**

  Phrase vs. Non-Phrase

- **Sequence structure**

  Patterns of content (C) and non-content (N) words, e.g., NNCN

- **Trial**
- **Manner, Place and Voicing**

  (of the first segment)

- **Transitional Probability**

  LogitAB = log(Frequency AB/((Frequency A*- Frequency AB)+1))

  LogitBC = log(Frequency BC/((Frequency B*- Frequency BC)+1))

  LogitCD = log(Frequency CD/((Frequency C*- Frequency CD)+1))

  LogitABC = log(Frequency ABC/((Frequency AB*- Frequency ABC)+1))

  LogitBCD = log(Frequency BCD/((Frequency BC* - Frequency BCD)+1))

  LogitABCD = log(Frequency ABCD/((Frequency ABC* - Frequency ABCD)+1))

Term	BNC	COCA
FreqA		
FreqB	t = 3.2	t = 4.2
FreqC		
FreqD		
FreqAB		
FreqBC		
FreqCD	t = -2.9	🟦
FreqABC	t = -5.4	🟦
FreqBCD		
FreqABCD		
LogitAB		
LogitBC		
LogitCD		
LogitABCD	🟦	t = -6.3
PhraseABCD(p)		
Length	t = 3.5	t = 3.9
NumSyll		
WordTypeABCD	F = 11.8	F = 14.9
Manner	F = 7.0	F = 2.4
Trial	t = 3.5	t = 3.5

# Subjects = 24   PhraseABCD(p) = phrases

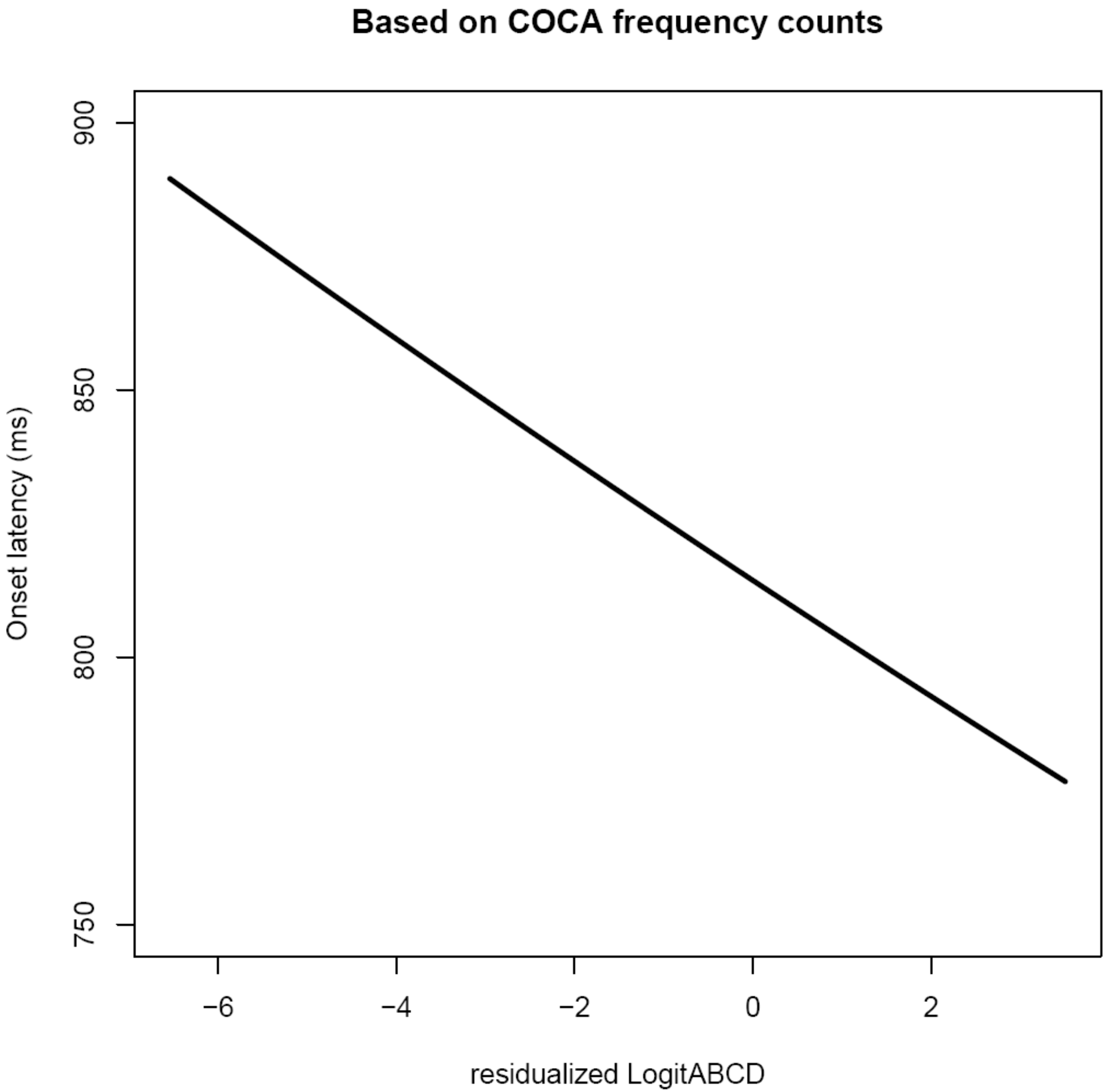
RESULTS ONSET LATENCY

quite a bit of overlap

in BNC results, FreqABC + FreqCD effect (all facilitatory)

in COCA LogitABCD effect (facilitatory)

negative is facilitatory
positive is inhibitory



Based on COCA frequency counts

RESULTS PRODUCTION DURATION

also quite a bit of overlap, but also some differences

Term	BNC	COCA
FreqA	🟦	t =-2.7
FreqB		
FreqC	t = -4.3	t =-3.7
FreqD		
FreqAB		
FreqBC		
FreqCD	t = -4.6	t =-2.2
FreqABC		
FreqBCD		
FreqABCD		
LogitAB	t = -3.7	🟦
LogitBC	t = -2.8	t = -3.5
LogitCD		
LogitABCD		
PhraseABCD(p)	t = -4.6	t = -2.5
Length	t = 15.2	t = 15.5
NumSyll	t = 17.7	t = 20.1
WordTypeABCD	F = 11.8	F = 16.5
Manner	F = 7.0	F = 10.4
Trial		
FreqC:PhraseABCD(p)	t = 3.8	🟦
FreqCD:PhraseABCD(p)	t = 2.5	🟦
LogitBC:PhraseABCD(p)	t = -2.1	🟦

FreqA in COCA, not BNC
LogitAB in BNC, not COCA

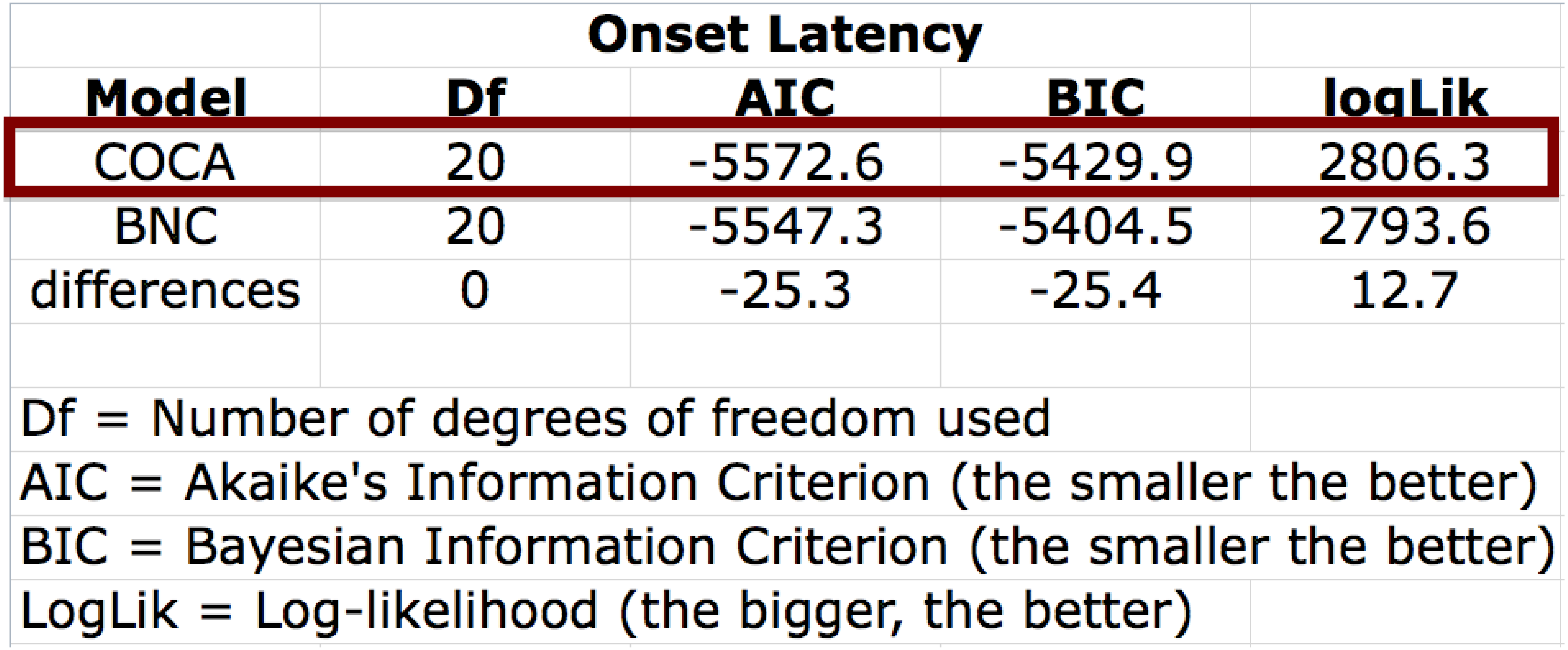
BNC has 3 interactions, COCA doesn’t

# Subjects = 17
PhraseABCD(p) = phrases

Results Summary

- Onset = retrieval of lexical knowledge + motor plan
  - overall chunk frequency facilitates onset latency, whereas freq of individual word inhibits it
  - The specific chunk differs between analyses.
- Duration = motor action
  - Bigram frequency/transitional probability do the work.
  - Some discrepancy between the exact chunks doing the work, but LogitBC and FreqCD in both BNC and COCA reduce production durations
  - Additional interactions with PhraseABCD in BNC

A More Formal Comparison of Models



	Onset Latency			
Model	Df	AIC	BIC	logLik
COCA	20	-5572.6	-5429.9	2806.3
BNC	20	-5547.3	-5404.5	2793.6
differences	0	-25.3	-25.4	12.7

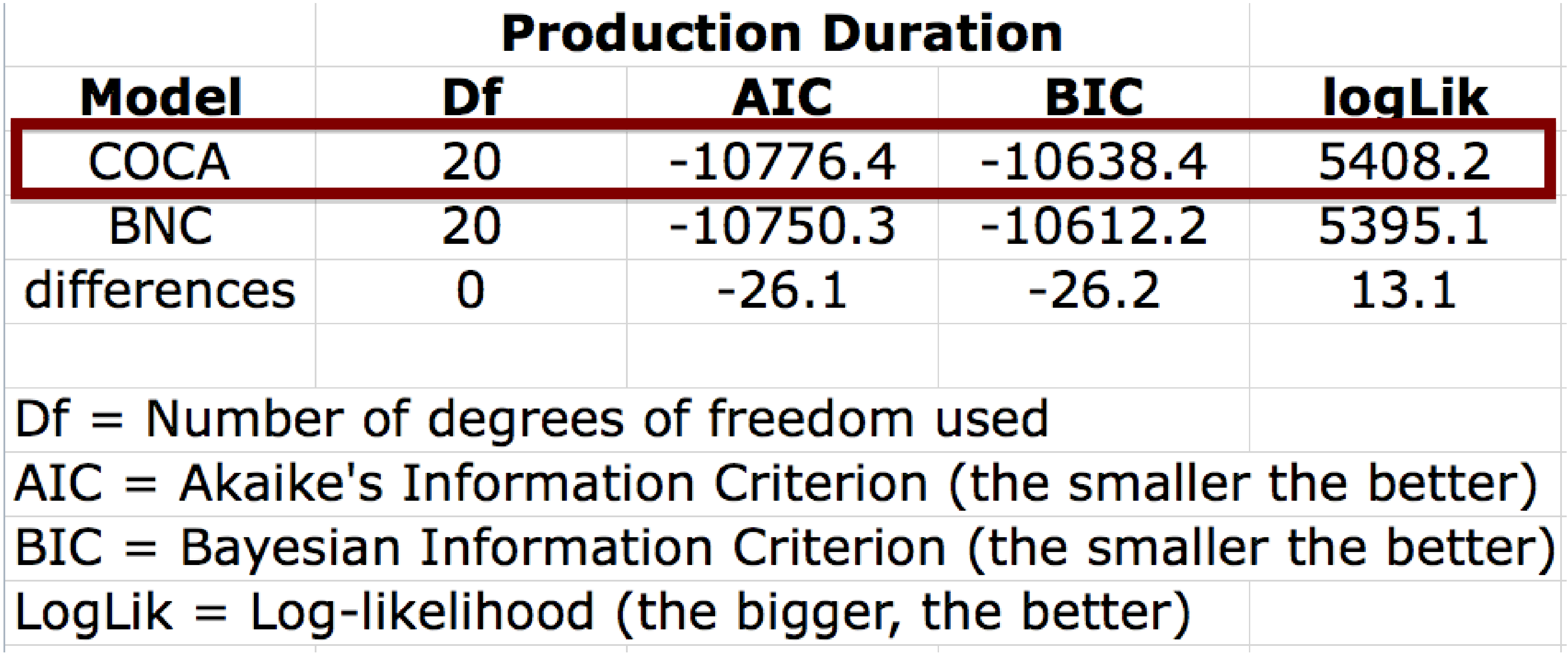
Df = Number of degrees of freedom used

AIC = Akaike's Information Criterion (the smaller the better)

BIC = Bayesian Information Criterion (the smaller the better)

LogLik = Log-likelihood (the bigger, the better)

A More Formal Comparison of Models



	Production Duration			
Model	**Df**	**AIC**	**BIC**	**logLik**
COCA	20	-10776.4	-10638.4	5408.2
BNC	20	-10750.3	-10612.2	5395.1
differences	0	-26.1	-26.2	13.1

Df = Number of degrees of freedom used

AIC = Akaike's Information Criterion (the smaller the better)

BIC = Bayesian Information Criterion (the smaller the better)

LogLik = Log-likelihood (the bigger, the better)

Discussion

- Perhaps the differences not only come from the differences between two English dialects, but is also from corpus size!
  - BNC = 100+ million words
  - COCA = 400+ million words
  - Frequency counts in COCA are thus more accurate than BNC
- Most likely a mix of dialect and corpus size, though this remains to be confirmed

Conclusions

- The general pattern of results from BNC and COCA are, to a large extent, comparable
- Some differences w.r.t. the exact variables doing the work
    - Not so good in Onset Latency analysis → Hoping to find a four-word frequency/probability effect and we did with COCA counts, but not with BNC counts.
- But if you’re just interested in chunk frequency effects in general, irrespective of which one exactly, well BNC or COCA should work fine.

Take home message

- Better to use a corpus that matches as closely as possible the dialect of the participants you will obtain data from.
- Most likely, the larger the corpus, the better the frequency counts will be
- If you can only obtain a corpus that doesn’t match the dialect of your participants, the results you’ll get shouldn’t be too far off.

Future work

- Compare British English speakers → analyze results with BNC and COCA
- Add ICE Canada to the comparisons
- Would ICE Canada better predict the results?
- Does size matter?
  - COCA 400+ million
  - BNC 100+ million
  - ICE Canada 1+ million