

# Measuring the formulaicity of language

---

Nick C. Ellis<sup>1</sup>, Ute Römer<sup>1</sup>, Matthew Brook O'Donnell<sup>1</sup>

<sup>1</sup>*University of Michigan*

{ncellis, uroemer, mbod}@umich.edu



Stefan Th. Gries<sup>2</sup>, Stefanie Wulff<sup>2</sup>

<sup>2</sup>*University of California, Santa Barbara*

{stgries, swulff}@linguistics.ucsb.edu



Colloquium "SLA and the inseparability of vocabulary and syntax"

Organised by Ute Römer and Stefanie Wulff

**AAAL 2009**

Denver Colorado - March 21-24, 2009

# Formulaic Language / Phraseology

---

- Growth of research into
  - Phraseology
  - Construction grammar
  - Native-like formulaic language use
  - The Idiom Principle
- In disciplines
  - Corpus Linguistics
  - Cognitive Linguistics
  - Psycholinguistics
  - Applied Linguistics
- Calls for
  - **clarity of conceptualization** across these fields, &
  - **basic investigation of metrics for operationalizations** of formulaicity in texts

Ellis, N. C. (2008). Phraseology: The periphery and the heart of language. In F. Meunier & S. Granger (Eds.), *Phraseology in language learning and teaching* (pp. 1-13). Amsterdam: John Benjamins.

Gries, S. T. (2008). Phraseology and linguistic theory: a brief survey. In S. Granger & F. Meunier (Eds.), *Phraseology: an interdisciplinary perspective*. Amsterdam: John Benjamins.

Römer, U. (Forthc.). Observations on the phraseology of academic writing: Local patterns – local meanings? In T. Herbst & S. Schueller (Eds.), *Chunks in the Description of Language. A Tribute to John Sinclair*. Berlin: Mouton de Gruyter.

Wulff, S. (2009). *Rethinking Idiomaticity. A Usage-based Approach*. London: Continuum.

# Project Goals

---

- Investigate how different measures of formulaicity, e.g.,
    - frequency of n-grams (2-9) above a threshold frequency level,
    - frequency of n-grams (2-9) above a threshold mutual information level,
    - frequency of phrase-frames,
    - frequency of reference n-grams from native corpora, etc.
  - Are affected by potential independent variables e.g.,
    - text length, mean length of utterance,
    - type- token ratio,
    - vocabulary frequency profiles, entropy, etc.
  - By potential text variables of interest, e.g.,
    - spoken/written register
    - Text-type (genre)
  - By potential subject variables, e.g.,
    - native vs. second language status
    - proficiency, etc.
  - By potential situational variables, e.g.,
    - degree of preparation / rehearsal
    - task working memory demands
-

# Presentation Goals

---

- Investigate how different measures of formulaicity, e.g.,
    - **frequency of n-grams (2-9) above a threshold frequency level,**
    - **frequency of n-grams (2-9) above a threshold mutual information level,**
    - frequency of phrase-frames,
    - **frequency of reference n-grams from native corpora, etc.**
  - Are affected by potential independent variables e.g.,
    - text length, mean length of utterance,
    - type- token ratio,
    - vocabulary frequency profiles, entropy, etc.
  - By potential text variables of interest, e.g.,
    - **spoken/written register**
    - **Text-type (genre)**
  - By potential subject variables, e.g.,
    - **native vs. second language status**
    - **proficiency, etc.**
  - By potential situational variables, e.g.,
    - degree of preparation / rehearsal
    - task working memory demands
-

# Measures of Formulaicity

---

# Why different measures of formulaicity?

---

## □ Frequency

- 'Lexical bundle' approach (Biber, Conrad et al. 2004; Biber, et al. 1999)
- Based solely on frequency - Methodologically straightforward
- Definition common in Applied and Corpus Linguistics
- But catches interesting ( 'on the other hand', 'it is possible') alongside uninteresting strings ('it has been' and 'and of the' )
- Many high frequency n-grams occur simply by dint of the high frequency of their component words, often grammatical functors

## □ Mutual Information

- MI statistical measure commonly used in the field of information science
  - Assess degree to which the words in a phrase occur together more frequently than would be expected by chance (Manning and Schütze 1999; Oakes 1998).
  - Higher MI score means a stronger association between the words, while a lower score indicates that their co-occurrence is more likely due to chance.
- 
- MI is a scale, not a test of significance

# Operationalizations and Metrics

---

## □ Different measures of formulaicity

- **Raw Frequency**: Type/token frequency of n-grams (2-9) above a threshold frequency level
  - WordSmith Tools (observe sentence boundaries)
  
- **Mutual Information**: Type/token frequency of n-grams (2-9) above a threshold mutual information level
  - Calculated MI for all 2- to 9-grams in the whole of BNCBaby occurring 12+ times
  - for each n we found the median MI, resulting in:
  
- **Reference List**: Type/token frequency of n-grams extracted from BNCBaby Academic

N	Median MI
2	2.234
3	6.723
4	13.085
5	20.835
6	38.925
7	53.612
8	69.046
9	79.962

# Experiment 1: Effects of Register & Genre

---

Spoken vs. written;  
Academic vs. non-academic

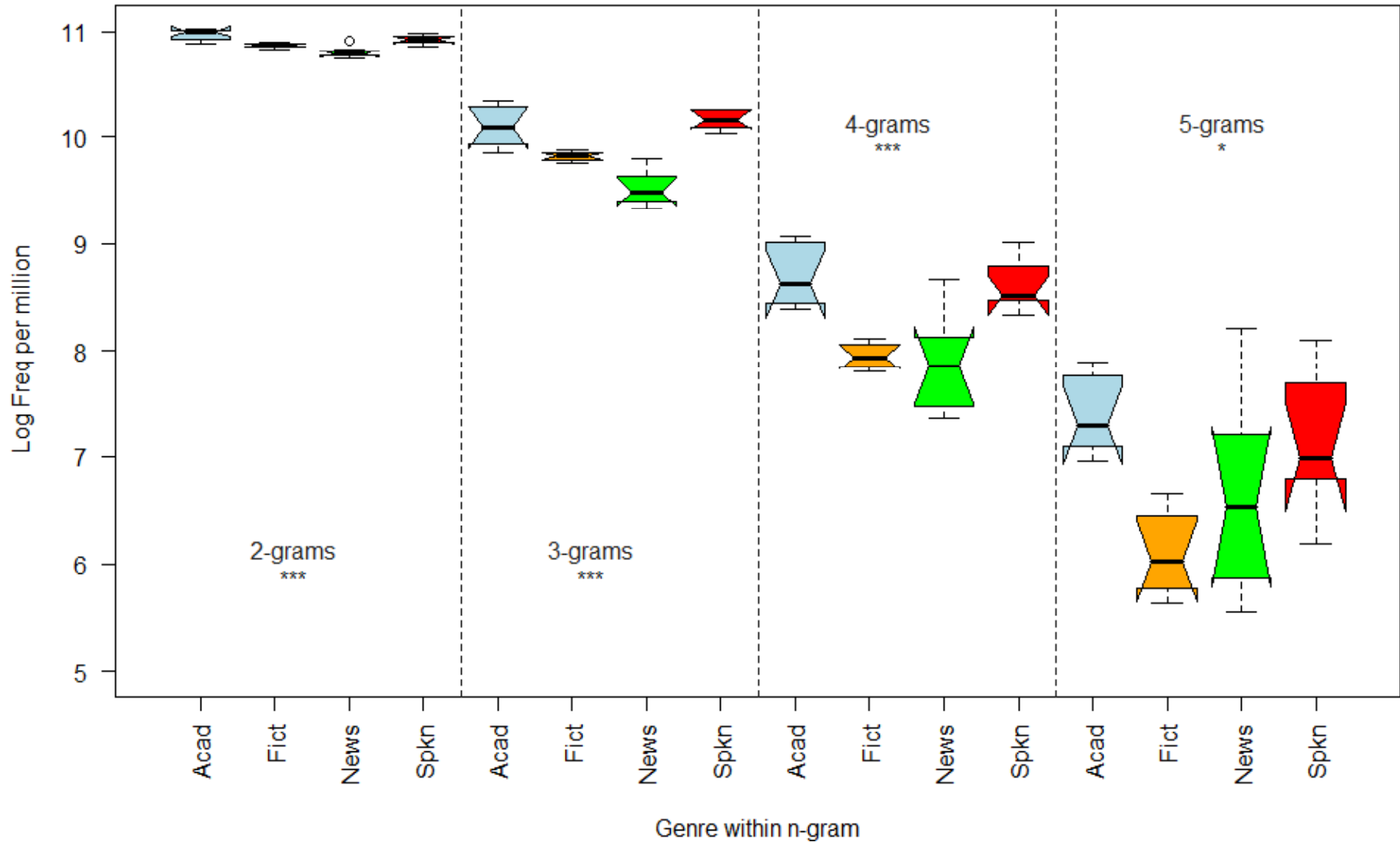
# Native language Reference Corpora: BNCBaby

---

- 1 million words each from four genres
  - Academic Writing (ACA) – 30 texts
  - Newspaper Texts (NEWS) – 97 texts
  - Fiction Excerpts (FIC) – 25 texts
  - Demographically Sampled **Spoken** Texts (DEM) – 30 texts
  
- Stratified into 8 samples of ~125,000 words each
  - A number partitioning algorithm used that attempts to balance token count and file count in each group
  
- Experimental design looking for significant differences across the 4 registers
  - Frequency threshold for 2- to 9-grams of 3 per ~125,000 words
  - Mutual Information (MI)

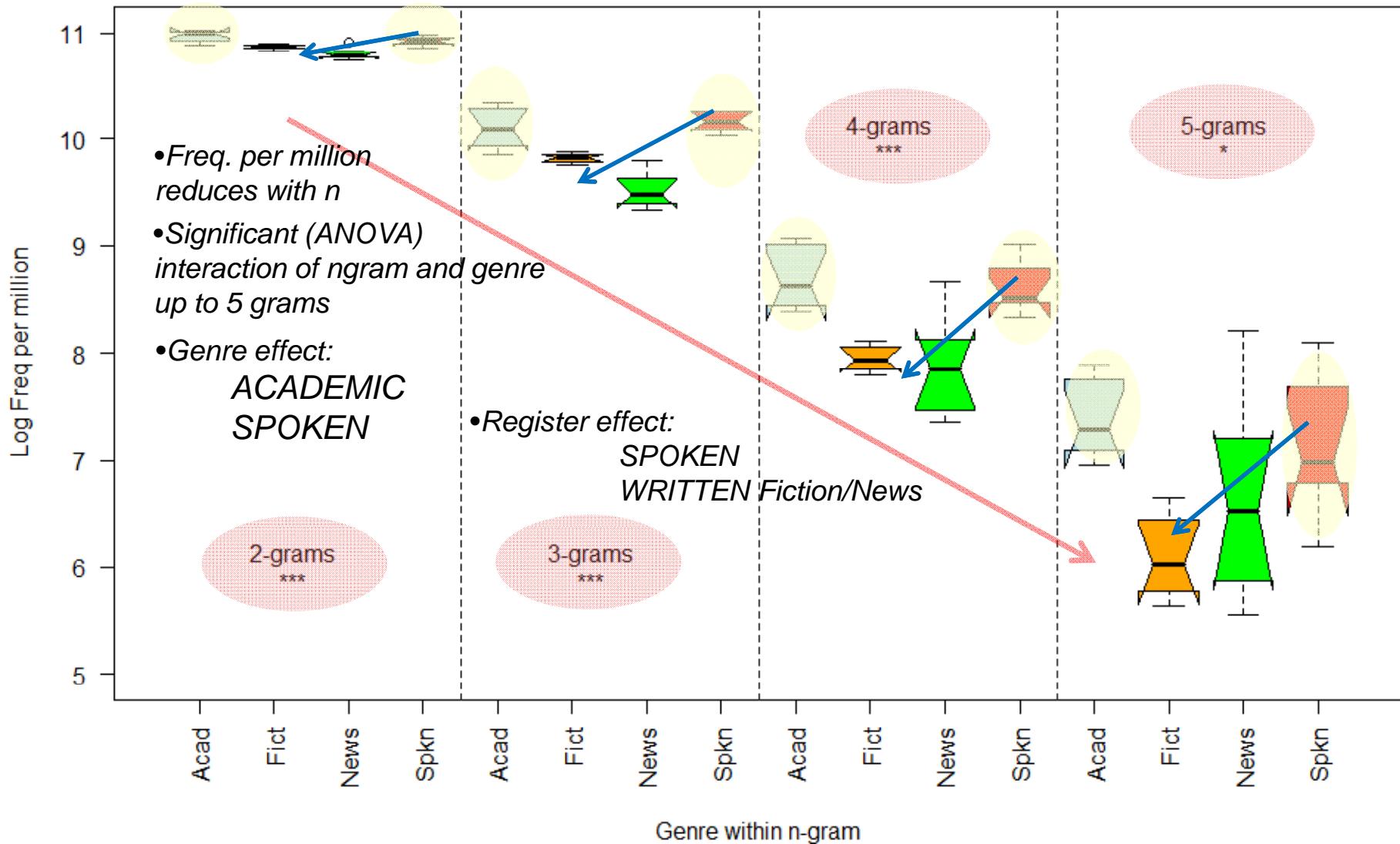
# Distribution of 2- to 5-grams across 8 groups samples for each text-type in BNCBaby (log frequency per million words)

raw frequency (2 to 5 grams occ. 3+)

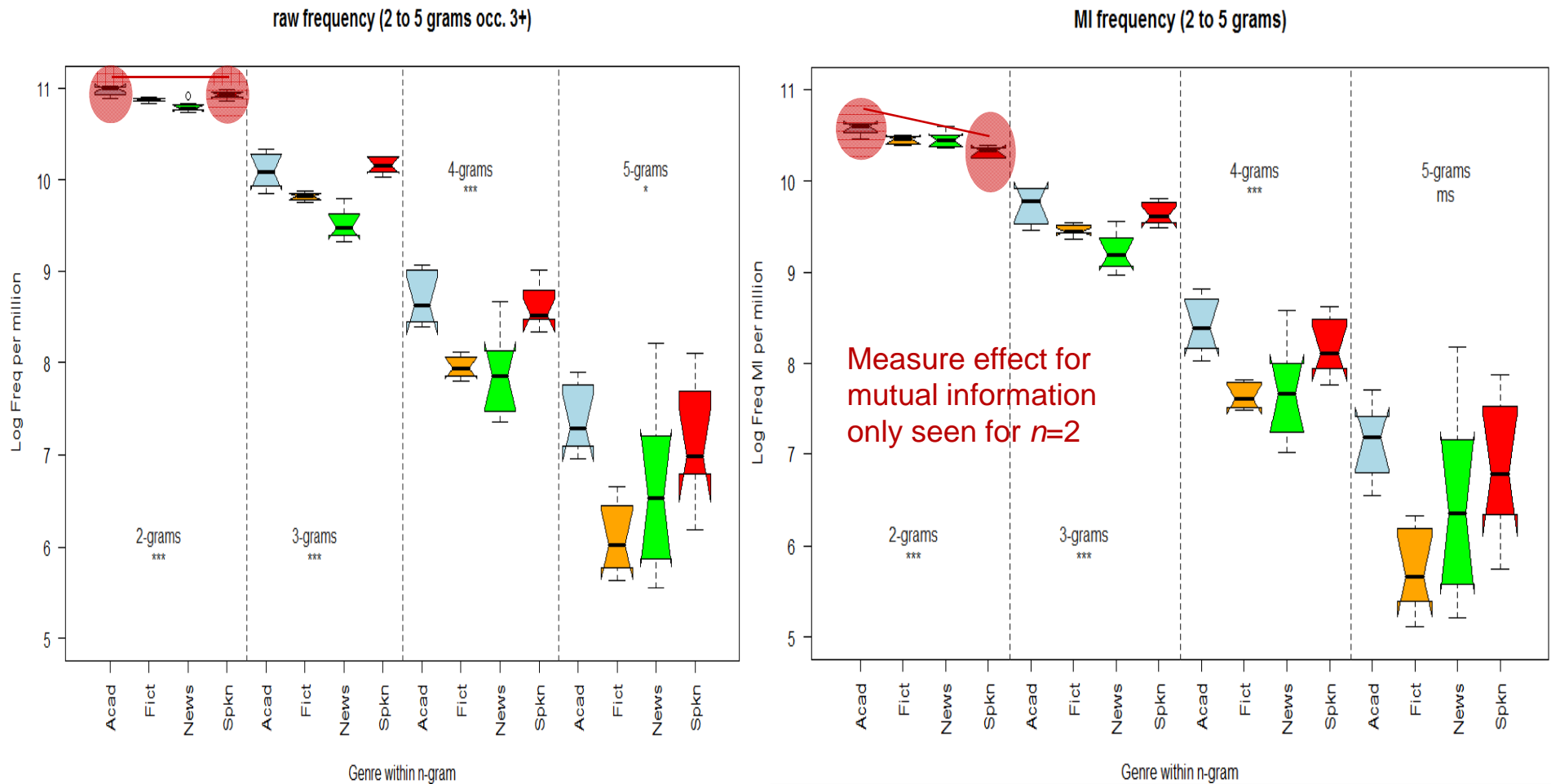


# Distribution of 2- to 5-grams across 8 groups samples for each text-type in BNCBaby (log frequency per million words)

raw frequency (2 to 5 grams occ. 3+)



# Distribution of 2- to 5-grams across 8 groups samples for each text-type in BNCBaby (log frequency per million words)



# Conclusions - Experiment 1

---

- Effects of **Register** – Spoken > Written Fiction or News (Frequency measure)
  
- Effects of **Genre** – Academic writing > Fiction or News (Frequency & MI)
  
- Effects of **Measure**
  - **Spoken language** is rich in high frequency formulas, but not so in 2-grams of high MI
  - At  $n > 2$  no such measure effects are found (MI and Frequency measure are similar)

# Experiment 2: Effects of SLA upon formula frequency

---

NS expert academic writing  
vs NS apprentices  
vs advanced NNS writing

# L1 Expert Academic Writing Corpora: BNCBaby Academic & Hyland

---

- BNCBaby Academic subset
    - 316,638 words
    - 9 files taken from the 30 files in the BNCBaby Academic section
  
  - Hyland subset
    - 369,653 words
    - 64 files taken from the 8 discipline categories of Hyland Corpus of Academic Writing (*Biochemistry, Electrical and Mechanical Engineering, Linguistics, Philosophy, Sociology and Physics*)
  
  - Stratified samples
    - BNCBaby Subset: Files selected on basis of token count ~ 38-40,000 words each, forming 8 groups
    - Hyland Subset: Stratified into 8 samples of ~ 46,000 words each
  
  - Experimental design providing L1 reference for ICLE subcorpora
    - Frequency threshold for 2- to 9-grams of 3 per ~40,000 words
    - Mutual Information
-

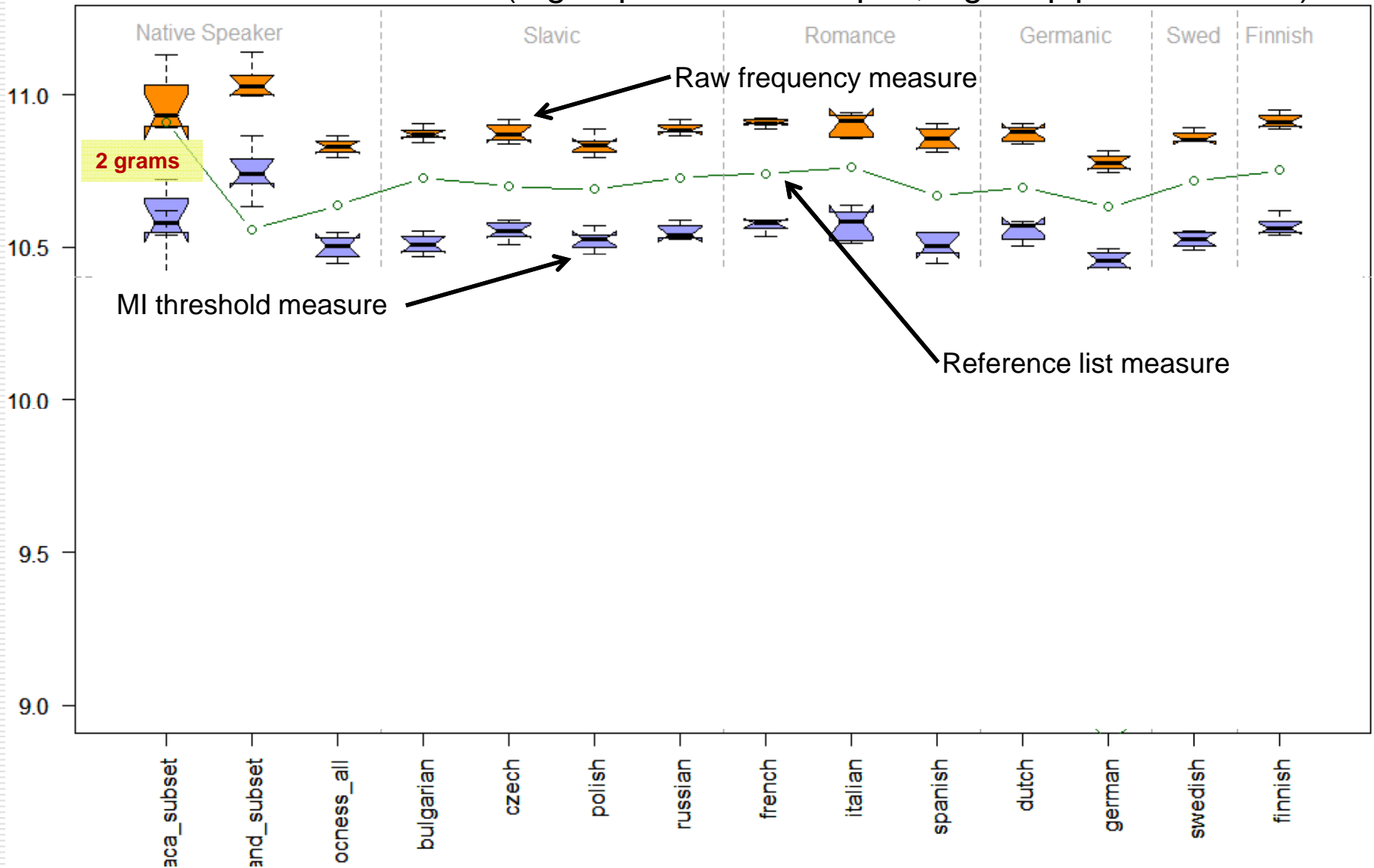
# L1 Student Reference Corpus

## LOCNESS & ICLE: Apprentice ACA Writing

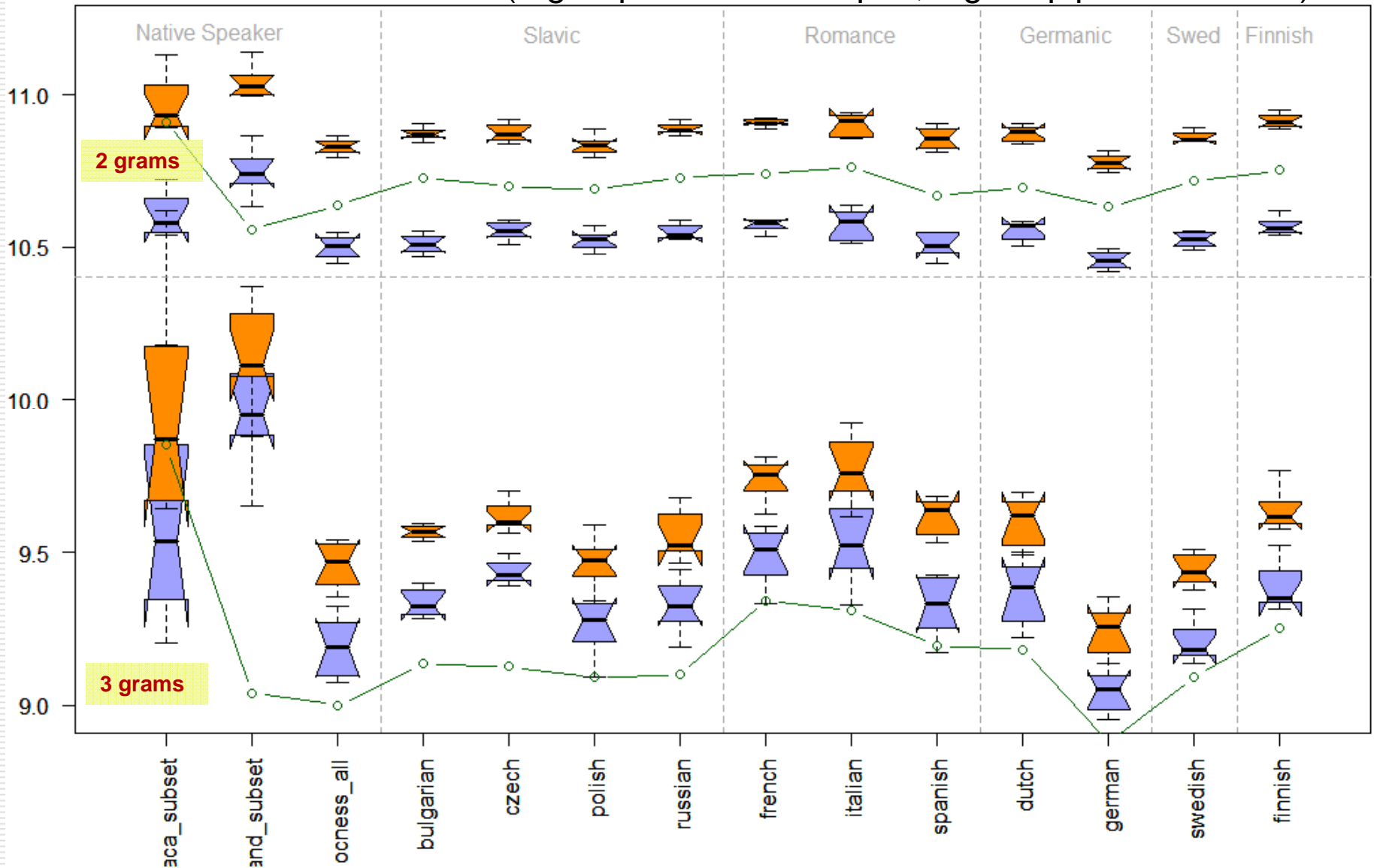
---

- LOCNESS - 269,839 words
    - British Undergraduates & British High School (A-level) students
    - American Undergraduates
  
  - ICLE – 11 L1 backgrounds, 200-300,000 words each
    - Bulgarian, Czech, Dutch, Finnish, French, German, Italian, Polish, Russian, Spanish, Swedish
  
  - Each stratified into 8 samples of ~30,000 words
  
  - Experimental design providing L1 reference for ICLE subcorpora
    - Frequency threshold for 2- to 9-grams of 3 per ~34,000 wds
    - Mutual Information
    - Comparison with BNCBaby Academic reference list
-

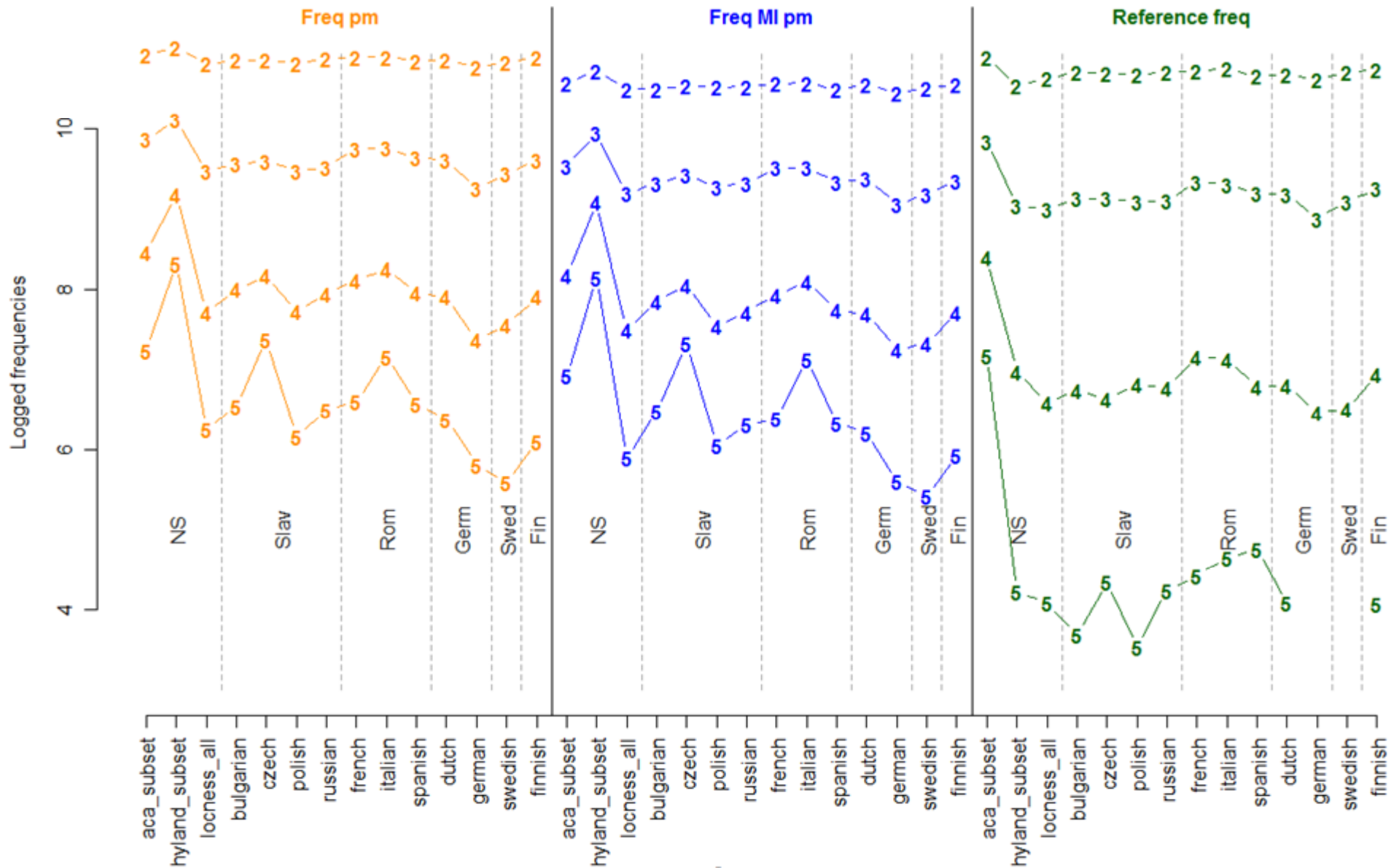
# Distribution of 2- & 3-grams across BNCCBaby Academic, Hyland, LOCNESS and ICLE corpora by frequency, MI and reference measures (8 groups for each corpus; log freq. per mil. words)



# Distribution of 2- & 3-grams across BNCCBaby Academic, Hyland, LOCNESS and ICLE corpora by frequency, MI and reference measures (8 groups for each corpus; log freq. per mil. words)



# Distribution of 2- to 5-grams across BNCBaby Academic, Hyland, LOCNESS and ICLE corpora by frequency, MI and reference measures (8 groups for each corpus; log freq. per mil. words)



# Conclusions - Experiment 2

---

- Native speaker expert Academic Writing is dense in formulaic use (both BNCBaby academic and Hyland corpus)
- Both NS (LOCNESS) and NNS (ICLE) apprentice writers use fewer formulas than Native expert writers
- We observe an **effect of proficiency** rather than nativeness (especially high numbers in Hyland)
- Variation between learners of different L1 backgrounds in ICLE
  - E.g. Italian >> German
- No strong effect of frequency and MI measures
- Reference measure:
  - NNS L1 groups less differentiated in terms of usage of NS academic formulas
  - Loss of Hyland peak (effect of text type and number)

# Case study: Effects of SLA upon formula functions

---

NS expert academic writing  
vs NS apprentices  
vs advanced NNS writing

# Functional classification of 4-grams

---

- Compiled lists from BNCCBaby Academic, Hyland, LOCNESS and ICLE subcorpora of 4-grams of 3+ frequency
  - Categorized all 4-grams in BNCCBaby Academic occurring 10+ times (290 altogether)
    - Content
    - Discourse organization
    - Evaluation/Stance
- Context (in concordance) taken into account
- Counted occurrences of these 4 grams (across function categories) in the other corpora

# Examples of 4-grams (see handout)

<b>Content</b>	<b>Discourse organization</b>	<b>Evaluation/Stance</b>
PER CENT OF THE	ON THE OTHER HAND	IT IS POSSIBLE TO
THE EXTENT TO WHICH	IN TERMS OF THE	IT IS IMPORTANT TO
IN THE UNITED STATES	IN THE CASE OF	THAT THERE IS A
A WIDE RANGE OF	THE END OF THE	IT IS CLEAR THAT
AT THE TIME OF	ON THE BASIS OF	ONE OF THE MOST
THE HOUSE OF COMMONS	AS A RESULT OF	CAN BE USED TO
THE REST OF THE	THE WAY IN WHICH	IT IS DIFFICULT TO
THE SIZE OF THE	AT THE END OF	IS LIKELY TO BE
THE NATURE OF THE	AT THE SAME TIME	THAT THERE IS NO
IN THE COURSE OF	IN THE FORM OF	MORE LIKELY TO BE

# Examples of 4-grams in context

another in the course of genetic evolution, the mode of  
 ory and, in the course of her employment there was injur  
 e acting in the course of his employment, but we must ta  
 In the course of negotiation. the Unionist lead  
 st event in the c In terms of the metric functions in (6.20), Panov  
 wareness in the c In terms of the metric functions of the line eleme  
 ination. In the c In terms of the metric functions of the Szekeres  
 esources in the chapter 6. In terms of the metric functions of the Szekeres  
 century; in the c provision in terms of the needs of users and to consider the  
 e abbey. In the c on for U in terms of a driver to realize that there is a considerable risk of an accid  
 fees. . In the c xpressed in terms of a the standard result that there is a high reliance on face validit  
 xpressed in terms of ngly taking the risk that there is a car travelling in the opposit  
 tropics in terms of . This is not to say that there is a " natural" style for you; all  
 variable in terms of curves it can be seen that there is a rapid change in log over a na  
 habitats in terms of a series of studies, that there is a link between genetically base  
 ure 15.1 in terms of eviews above suggest that there is a need for improvements in user  
 cumented in terms of work, have suggested that there is a systematic bias in favour of  
 xplained in terms of ratures and suggests that there is a relaxation process active in  
 itive), she suggests that there is a hidden argument which takes t  
 scale. This suggests that there is a strong argument for having tw  
 It is true, I think, that there is a certain sort of eclecticism,  
 orality. People think that there is a special barrier here to the a

content-related

discourse  
organization

evaluation/stance

# Functional classification of 4-grams



Figure based on **relative** frequencies

# Functional classification of 4-grams

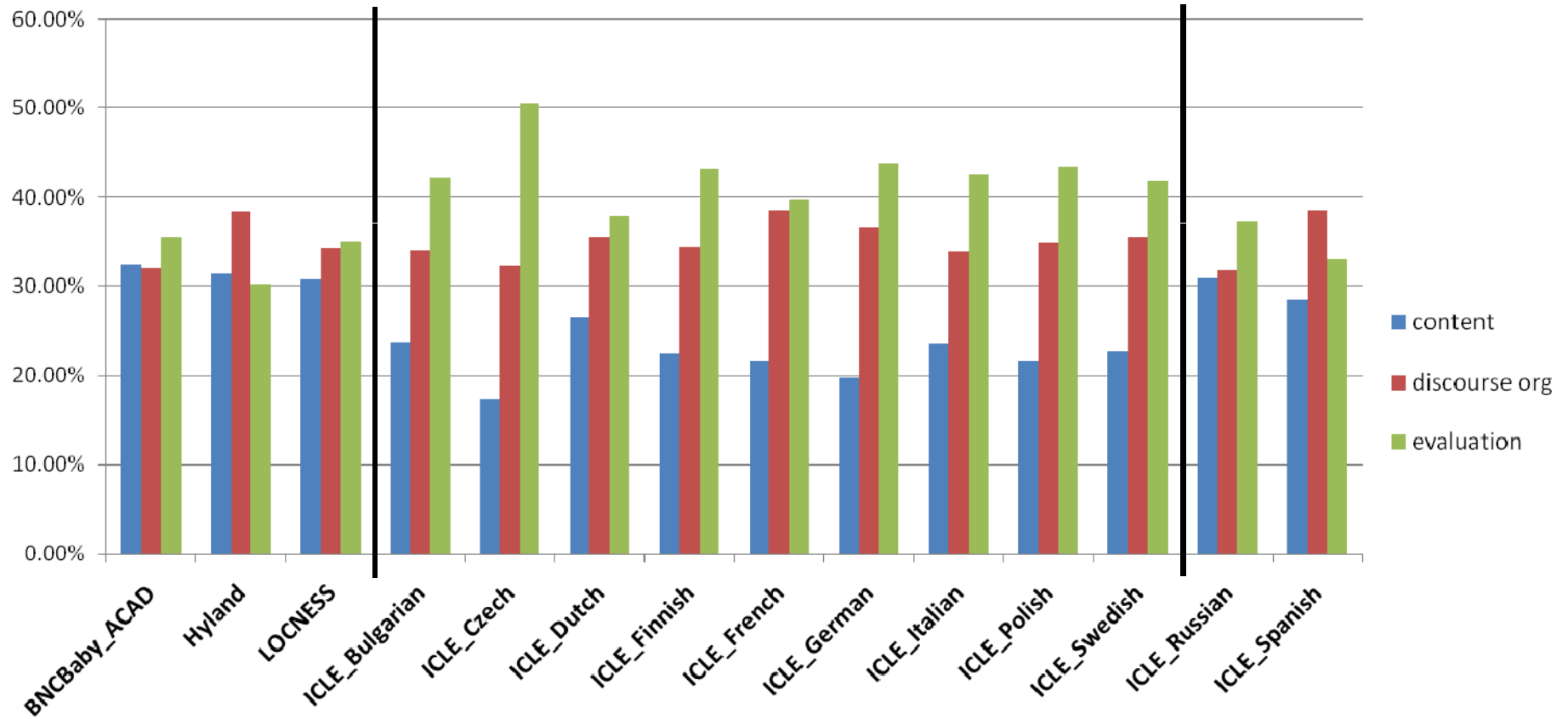


Figure based on **relative** frequencies

## Conclusions & observations – Case study

---

- Most frequent 4-grams in NS academic writing are roughly evenly distributed across three **functional categories**, both in apprentice and expert writing
- NNS learner-universal (9 out of 11 ICLE subsets) deviations from NS norm:
  - Lower shares of content-related (not topic-specific) 4-grams
  - Similar shares of discourse organizing 4-grams
  - Higher shares of 4-grams that express evaluation
- Nativeness seems to have an effect on functional distribution (different from frequency/MI findings)



# Conclusions & Outlook

---

# Conclusions

---

- ❑ We have observed effects of **register**, **genre** and **measure** on the distribution of 2- to 5-grams in BNCBaby.
- ❑ In both **apprentice** and learner academic writing we found fewer n-grams than in **expert** academic writing. Apprentice and learner writers are remarkably similar in their usage in all measures used. → **proficiency effect**
- ❑ **Learner L1** impacts n-gram usage, possibly along language family lines.
- ❑ There is a consistent adoption of general academic formulas (2- to 5-grams) by all ICLE writers. But there are notable differences with regard to the way in which these formulas (4-grams) are **functionally distributed** compared to NS writers. → **nativeness effect**

# Future goals

---

- Investigate how different measures of formulaicity, e.g.,
  - frequency n-grams (2-9) above a threshold frequency level,
  - frequency of n-grams (2-9) above a threshold mutual information level,
  - frequency of phrase-frames,
  - frequency of reference n-grams from native corpora, etc.
- Are affected by potential independent variables e.g.,
  - text length, mean length of utterance,
  - type- token ratio,
  - vocabulary frequency profiles, entropy, etc.
- By potential text variables of interest, e.g.,
  - spoken/written genre
- By potential subject variables, e.g.,
  - native vs. second language status... CHILDES & ESF SLA (longitudinal)
  - Proficiency (add MICUSP/BAWE), individual differences, etc....
- By potential situational variables, e.g.,
  - degree of preparation / rehearsal
  - task working memory demands