A Corpus-based Analysis of Lecturing Language
in Face-to-face Courses and MOOCs

Zhi Li

March 26, 2018

AAAL 2018, Chicago, IL

Research Background

- Research motivation
  - Massive Open Online Courses or MOOCs (Hew & Cheung, 2014; Margaryan et al., 2017)
  - (continued) Domain analysis for an English language proficiency test for academic purposes

- Importance of this study
  - Test development and validation (Cushing, 2017)
  - Material development for English for academic purposes (EAP) courses (Flowerdew, 2013)

Research Background – Lecturing Language

- Lecturing language as a genre
  - Similar terms: Teacher talk, Classroom discourse, **Instructional discourse**
  - Important roles of lecturing language
  - Features of lecturing language
    - Informational & interactive language (Biber, 2006)
    - Online production (time constraints) vs. Prepared scripts (Biber et al., 2002)
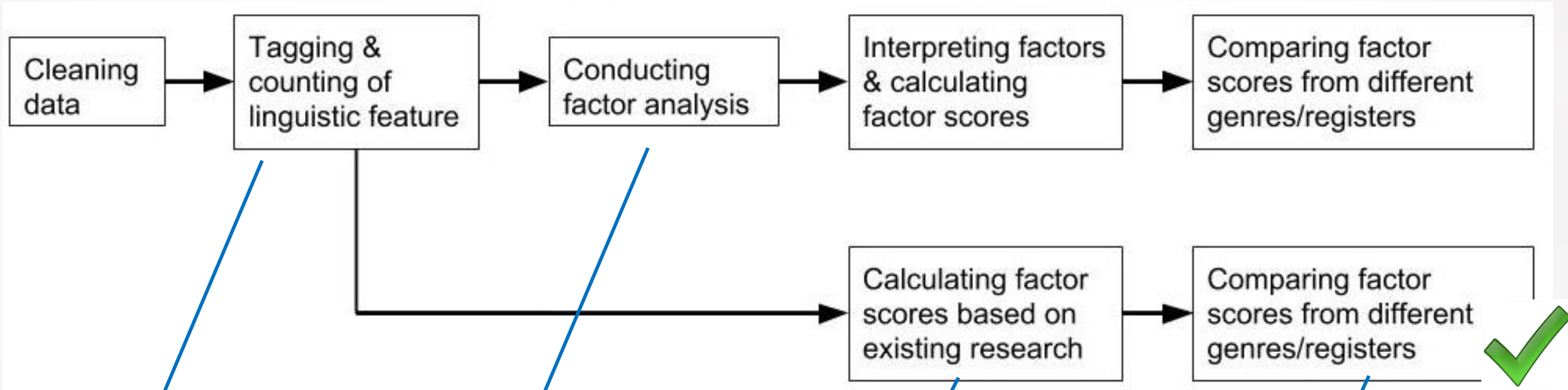   - Oral-literate continuum (Biber 2006; Csomay, 2006)

Research Background – Lecturing Language

- Situational factors and lecturing language

  - Class size (Lee, 2009)
  - Level of education (Barbieri, 2015)
  - Instructors' first language (Fung & Carter 2007)
  - **Delivery mode (face-to-face courses and online courses) ***

Research Background – Analytical Approaches

- Corpus-based analyses
  - Individual linguistic features
    - Lexical bundles (Csomay, 2012),
    - Discourse markers (Moreno et al., 2006)

  - **Multidimensional analysis or MDA**
    - Registerial variations (e.g., spoken vs. Written registers) (Biber, 1988)
   - World Englishes (Xiao, 2009)

Multidimensional Analysis (MDA)

Cleaning data → Tagging & counting of linguistic feature → Conducting factor analysis → Interpreting factors & calculating factor scores → Comparing factor scores from different genres/registers

Tagging & counting of linguistic feature → Calculating factor scores based on existing research → Comparing factor scores from different genres/registers ✓

Types of taggers:
Biber's tagger,
CLAWS POS tagger,
Nini's tagger

Number of factors or dimension,
Factor loading for individual variables

Factor scores for each text on all factors
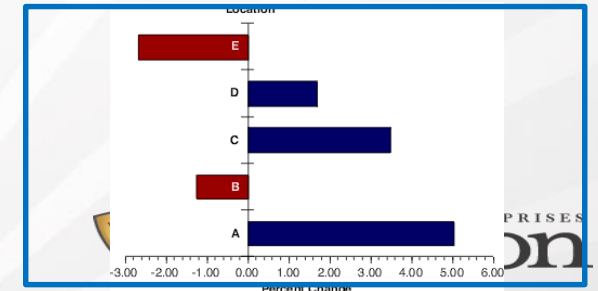
Research Background – Analytical Approaches

- **Biber's (1988) 6-dimensions**
  - 67 linguistic features from Biber's tagger (mostly lexico-grammatical features)
  - Based on 481 spoken and written texts in 23 major register areas (e.g., academic prose, prepared speeches, conversations)
  - From the Lancaster-Oslo-Bergen Corpus and the Lond-Lund Corpus

- Applications of Biber's 6-dimensions
  - Registerial variations (e.g., spoken vs. Written registers) (Biber, 1988)
  - University registers (Biber et al., 2012)
  - Natural Conversations and TV shows (Al-Surmi, 2012)
  - Forensic linguistic analysis/authorship identification (Nini & Grant,,2013)
  - Teaching ESL students (Aguado-Jimenez et al., 2012)

MDA – Biber's (1988) 6- Dimensions

Dimensions	Example features
D1 Involved vs. Informational Production	**+** private verbs, *that*-deletion, *do*-as pro-verb, 1st & 2nd personal prons. **-** type/token ratio, preposition, nouns, word length
D2 Narrative vs. Non-narrative Concerns	**+** past tense verbs, 3rd person prons., perfect aspect verbs
D3 Elaborated vs. Situation-dependent References	**+** *WH*-relative clauses in object positions, nominalizations; **-** time adverbials, place adverbials, adverbs
D4 Overt Expression of Persuasion	**+** infinitives, prediction modals, suasive verbs, conditional subordination
D5 Impersonal vs. Non-impersonal Style	**+** conjuncts; agentless passives, past participial adverbial clauses
D6 Online Informational Elaboration	**+** *that* clauses as verb complements, demonstratives

Research Design – The Corpora

- the Michigan Corpus of Academic Spoken English (MICASE)
  - Collected at the University of Michigan from 1997 to 2001
  - Total running words of 1.8 million (Simpson et al., 2002)
    - 152 academic speech samples in 15 categories such as lectures, colloquia, etc.
  - 28 highly or mostly monologic lectures

- The MOOC corpus
  - 28 course offered by universities in North America on Coursera
    - Transcripts available online
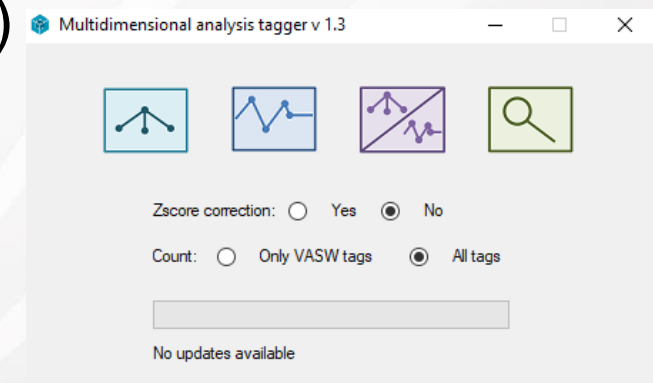 - A variety of visuals used with no presence of audience

Research Design – The Corpora
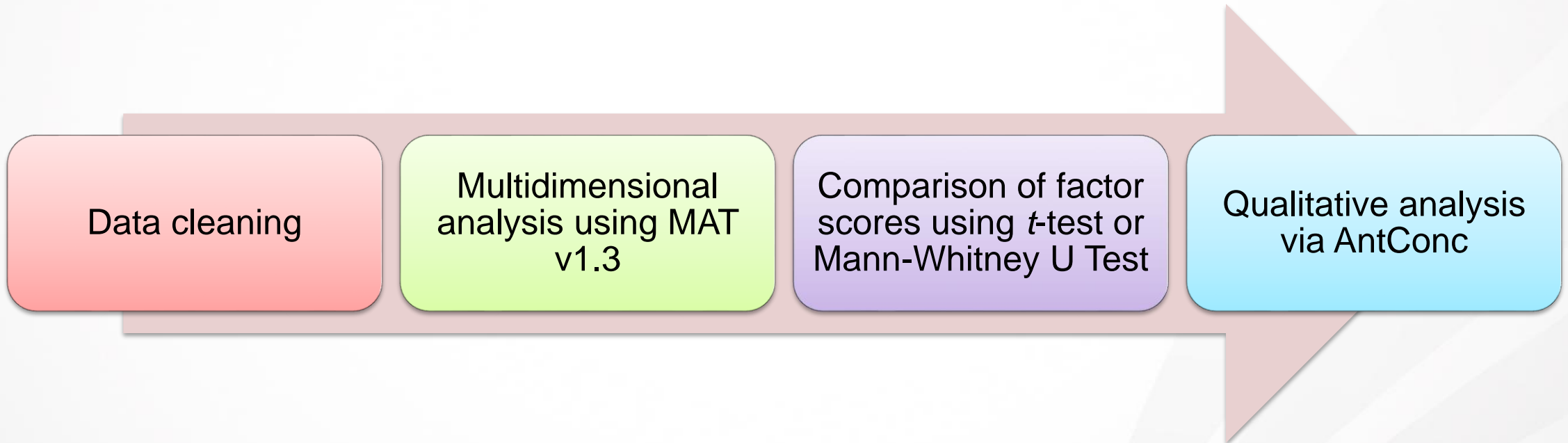
- Overview of the corpora

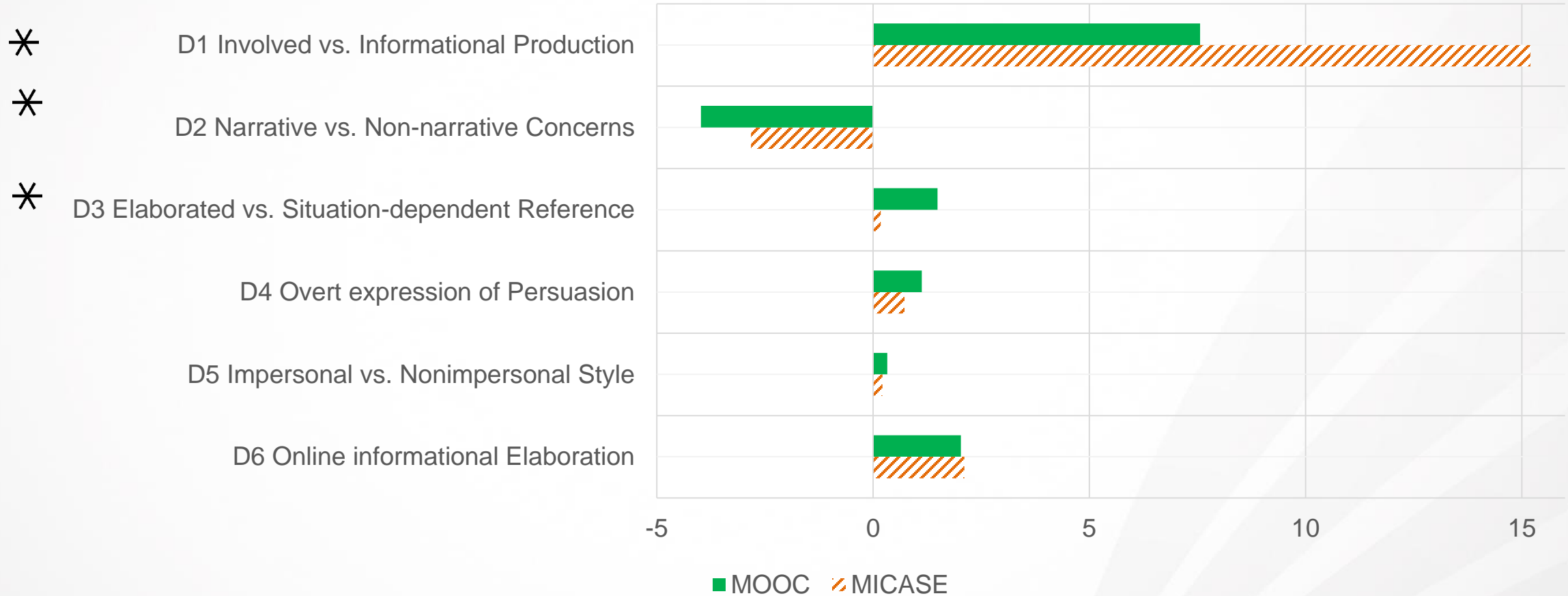| Corpus	# of scripts	Min/Max length	Mean length	S.D. of length	Total word count	Disciplines
MOOC	28	4,146/16,214	9,084.3	3,017.1	254,359	Soft sciences: 15 Hard sciences: 13
MICASE	28	3,798/14,571	9,092.5	2,798.8	254,590	Soft sciences: 18 Hard sciences: 10

Research Design – Analytical Tool

- Multidimensional analysis tagger (MAT) v 1.3 (Nini, 2015)
  - Tagged features
    - 67 lexico-grammatical features (e.g., past tense verbs, that clause as verb complements)
    - Reliability of the tool (Nini, 2015)
  - Biber's (1988) 6-dimensions
    - Biber et al. (2002); Csomay (2006); Reppen (2006)

Multidimensional analysis tagger v 1.3

Zscore correction: ○ Yes ◉ No

Count: ○ Only VASW tags ◉ All tags

No updates available

Research Design – Data analysis

Data cleaning → Multidimensional analysis using MAT v1.3 → Comparison of factor scores using *t*-test or Mann-Whitney U Test → Qualitative analysis via AntConc

Results – Overview of the 6 Dimensions

Results – D1 Involved vs. Informational Production

- MICASE
  - *this is ON your handout, and here's a, a caterpillar a pine looper, and you can **see** we have these fluctuations but it turns out they are, IN a a pattern. (LEL175JU112)*
  - *so I wanted to address that since some OF you had experienced it. now we'll turn the lights back on and talk ABOUT loops. (LEL295JU035)*
- MOOC
  - *Also, IN this expression, you might **recognize** this FROM the previous class, as describing a probability distribution. (BS01)*
  - *Taking a cue FROM Maxine Greene and her notion that aesthetic experiences create openings FOR young people and others to **see** their lives FROM new perspective, what does this **mean** FOR society? (MU01)*

D1 Involved vs. Informational Production

-5   -3   -1   1   3   5   7   9   11   13   15

■ MOOC ■ MICASE

D1 Involved vs. Informational Production
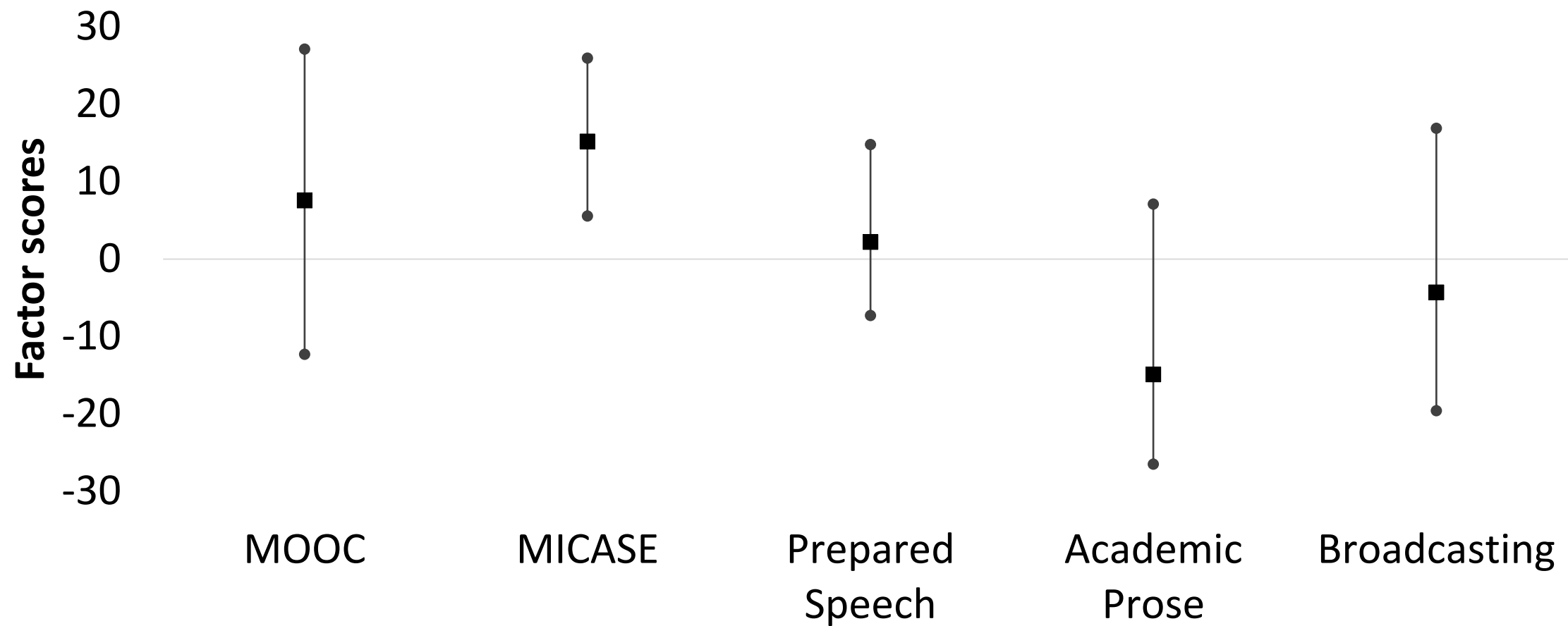
Results – D2 Narrative vs. Non-narrative Concerns

- MICASE
  - *But Kampei because <u>he</u> DIED, because of the way that <u>he</u> DIED <u>he</u> MADE a great sacrifice for <u>his</u> lord <u>he</u> finally KILLED <u>himself</u>. (LEL140SU074)*
  - *I WENT over that Armstrong Neuendorf and Brentar article in class where <u>they</u> LOOKED at entertainment versus news. (LEL220SU073).*

- MOOC
  - *The Romans BELONGED to a group of people who HAD made settlements around the area of Rome at a number of sites. (RAA01)*

D2 Narrative vs. Non-narrative Concerns

| -5 | -3 | -1 | 1 | 3 | 5 | 7 | 9 | 11 | 13 | 15 |

■ MOOC ■ MICASE

# Results – D3 Elaborated vs. Situation-dependent References

- MICASE
  - *these had been torn down by Sulla seventeen years **earlier**, and and Caesar's fixed them back up. on one occasion <u>when</u> he goes after a political opponent in sixty-three B-C. (LEL215SU150)*

- MOOC
  - *This goddess <u>whose</u> name, Calypso, comes from a verb meaning to hide. (GM01)*
  - *We looked at several VARIATIONS that consider ASSUMPTIONS about future growth, and we came up with a number of formulas. (FFE01)*

D3 Explicit vs. Situation-dependent reference
-5   -3   -1   1   3   5   7   9   11   13   15
MOOC   MICASE
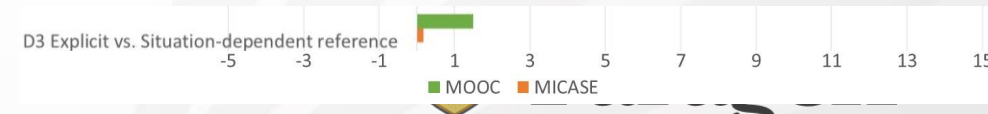
Results – D4 Overt Expression of Persuasion

- MICASE
  - *the idea behind an I-Q test is you've got somebody really smart, they <u>should</u> be going out and doing smart things right. (LES500SU102)*
- MOOC
  - *Every measurement <u>must</u> have two things. It <u>must</u> have a number and a unit. (CHE01)*
  - *This WILL ALSO BE true **if** an improvement in technology stimulates investment spending. (POM01)*

D4 Overt expression of persuasion

-5   -3   -1   1   3   5   7   9   11   13   15

■ MOOC   ■ MICASE
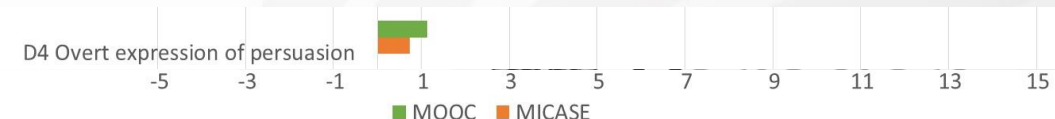
Results – D5 Impersonal vs. Non-impersonal Style

- MICASE
  - *and then it's FURTHERMORE **obscured** at the same time, it's **highlighted**, by, the placement of this light fixture which ON THE ONE HAND literally like bars your vision. (LEL320JU147)*
- MOOC
  - *Based on the coupon <u>divided</u> by the discount or premium price. So let's suppose that coupon rate of three and a half percent is being **paid** when the bond price is $700. (FFE01)*
  - *But they use randomization NONETHELESS, just in a different way and for a different purpose. (SPN01)*

D5 Abstract vs. Non-abstract information

-5   -3   -1   1   3   5   7   9   11   13   15

MOOC   MICASE

Results – D6 Online Informational Elaboration

- MICASE
  - *therefore, THIS was the kind of context <u>that</u> Haeckel was writing in, in the late nineteenth century he had THOSE questions in mind and he was thinking about THOSE questions. (LEL115SU107)*

- MOOC
  - *And THIS is the same type of bone <u>that forms </u>in the inside of the long bones of your arms and legs. (PEV01)*
  - *And to understand their roles, let's look at, for example, THIS equation right there. So, THIS applies to any action of player 1. (GT1)*

D6 Online informational elaboration

-5   -3   -1   1   3   5   7   9   11   13   15

■ MOOC ■ MICASE

Results – Summary

- Impact of course delivery modes on instructional language
  - Differences in
    - Interactivity and information density (D1)
    - Degree of narrativeness (D2)
    - Explicitness in references (D3)
  - Similarities in
    - Use of persuasive language (D4)
    - Impersonal language (D5)
    - Elaboration of information (D6)

Implications – EAP Teaching & Testing

- Modeling authentic lecturing language
  - What to look for in material development

  - What to teach in EAP courses

Limitations and Future Studies

- The corpora data
  - Samples from MICASE as a reference corpus
    - levels of interactivity, single institution
  - Lack of multimodal elements in both corpora

- The analytical approach (MDA)
  - Biber's (1988) dimensions vs. new genre-specific dimensions

Future Studies

- Future studies
  - New multidimensional analysis with larger corpora data (and a distinction of disciplines)
  - Analysis of specific linguistic features of lecturing language
  - Multimodal analysis of lecturing language
  - Learners' responses to the variations of lecturing language

Selected references

- Biber, D. (1988). *Variation across speech and writing*. Cambridge: Cambridge University Press.
- Biber, D., Conrad, S. M., Reppen, R., Byrd, P., & Helt, M. (2002). Speaking and writing in the university: A multidimensional comparison. *TESOL Quarterly, 36*(1), 9–48.
- Csomay, E. (2005). Linguistic variation within university classroom talk: A corpus-based perspective. *Linguistics and Education*, 15(3), 243–274.
- Csomay, E. (2012). Lexical bundles in discourse structure: A corpus-based study of classroom discourse. *Applied Linguistics, 34*(3), 1–21.
- Flowerdew, L. (2013). Corpus-based research and pedagogy in EAP: From lexis to genre. *Language Teaching, 48*(1), 1–18.
- Morell, T. (2004). Interactive lecture discourse for university EFL students. *English for Specific Purposes, 23*(3), 325–338.
- Moreno, C. P. A., Chambers, A., & O'Riordan, S. (2006). Integrating a corpus of classroom discourse in language teacher education: The case of discourse markers. *ReCALL, 18*(1), 83–104.
- Nini, A. (2015). Multidimensional analysis tagger 1.3. Retrieved from http://sites.google.com/site/multidimensionaltagger
- Rampton, B., Roberts, C., Leung, C., & Harris, R. (2002). Methodology in the analysis of classroom discourse. *Applied Linguistics*, 23(3), 373–392.

TESTING ENTERPRISES
Paragon

Thank You!

Questions & Comments?

Zhi Li (zli@paragontesting.ca)

Paragon Testing Enterprises, Inc. Vancouver, Canada