

March 7 (or 8!), 2022

ON THE PERILS OF OPAQUE MEASURES AND METHODS: Toward increased transparency

Tove Larsson

Northern Arizona University

Tove.Larsson@nau.edu



Greetings from Arizona!

OUTLINE

- Setting the scene: Description vs. prediction in CL
- Linguistically interpretable vs. opaque measures in the field
- Does it matter? Case study: syntactic complexity
- Suggested ways forward

LINGUISTIC DESCRIPTION VS. PREDICTION IN CORPUS LINGUISTICS

SETTING THE SCENE

- Linguistic description vs. predictive studies (Biber et al., 2020)
- Fundamentally different goals and methods
- Problematic if
 - there is a methodological mismatch: the method of one is used for the goal of the other
 - the goals are not identified as different



GOAL AND FOCUS

Linguistic description

- describing **the characteristics of groups** (and in the process, distinguishing among the groups)
- “The primary focus of descriptive studies is **the linguistic interpretation** of specific [linguistic] characteristics.”

Predictive studies

- to **predict group differences** (e.g., proficiency levels)
- “give **little or no attention to the linguistic interpretation** of omnibus measures, because such measures are not designed to be linguistically interpretable.”
- If it works, it works!

Biber et al. (2020: 3)

HOW DOES IT AFFECT OUR CHOICE OF MEASURES?

Stance adverbs (e.g., *interestingly*, *clearly*)

- Most often studied in their own right (linguistic description)

T-units

- Not often studied in its own right (~~linguistic description~~)
- Instead used as a means of studying/measuring something else (predictive studies)

LINGUISTICALLY INTERPRETABLE VS. OPAQUE MEASURES

LINGUISTICALLY INTERPRETABLE MEASURES

- A measure is interpretable “when its scale and values represent a **real-world language phenomenon that can be understood and explained.**”

Requirement:

- all variables have **clear operational definitions**
 - e.g., “relative clauses” may seem straightforward, but is it?
 - Only finite relative clauses (e.g., *the construction that was analyzed in the 2007 study*) or both finite and non-finite relative clauses (e.g., *the finding discussed in the 2008 study*)?
 - Larsson et al. (in preparation): main clauses??

Egbert, Larsson, & Biber (2020)

OPAQUE MEASURES

- \neq linguistically interpretable measures
- when its scale and values
 - **confound phenomena** or
 - DO NOT represent a **real-world language phenomenon that can be understood and explained**

Linguistically interpretable

Opaque



Frequency of adverbs

Means

Complex nominals

2, 3, 2
 $\bar{x} = 2.333$

WHERE DO WE GET MEASURES FROM?

- The large amounts of data typical of most empirical corpus linguistics studies necessitate computational tools to help process them
- We can either
 - use existing software tools
 - develop our own programs
- The field tends to rely heavily on pre-existing software tools

WHERE DO WE GET MEASURES FROM?

- These tools thus have a strong influence on current research practices in quantitative corpus linguistics
- → it is of the utmost importance that we critically examine the results they provide



TOWARD INCREASED ACCURACY AND TRANSPARENCY

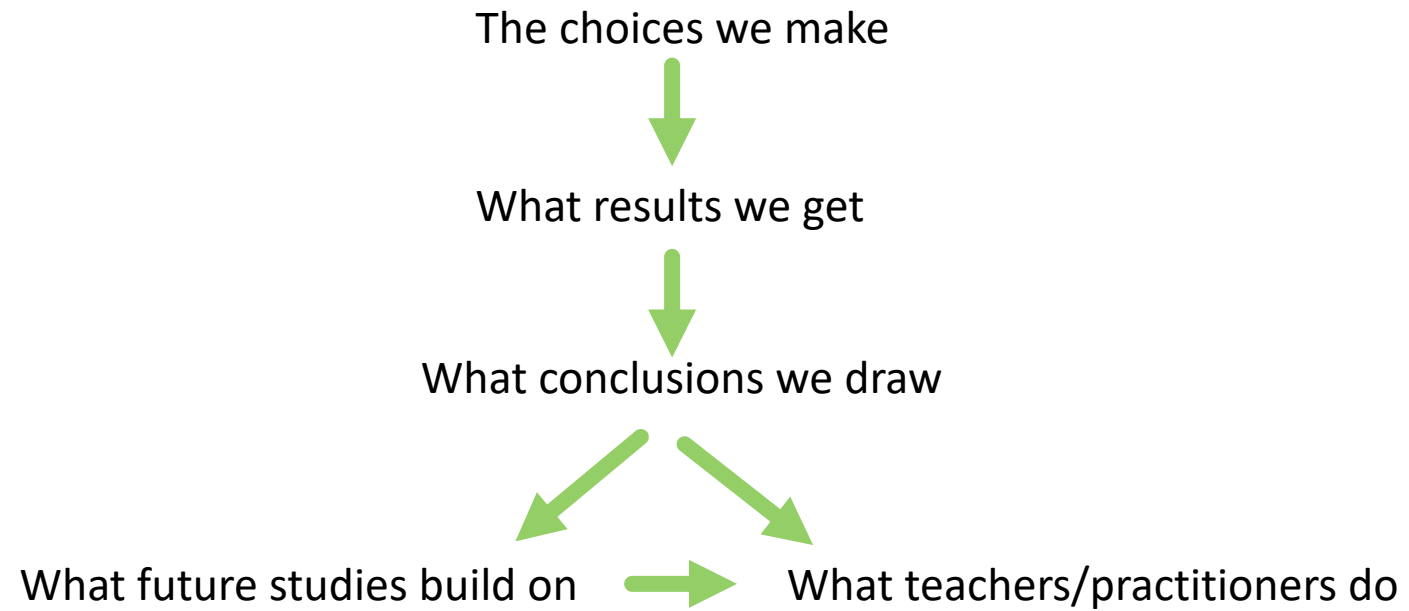
- It is difficult, if not impossible, for any tool to achieve perfect accuracy
 - not just accept results at face value
 - make sure to test (and report) the precision and recall
- Errors are not likely to be distributed randomly across all features
- Even more important when tools are applied to data that they were not developed for
 - e.g., most taggers and parsers are developed for and trained on native-speaker Standard English data – L2 data?? (Picoral et al., 2021)

TOWARD LINGUISTIC INTERPRETABILITY

- Accuracy is a threat to the validity of our results, but...
- “Perhaps the most serious risk to researchers using available tools is that many of the quantitative measures provided by corpus analysis software **do not have transparent linguistic interpretations**.
- In some cases, these are measures that
 - have **no direct counterparts in linguistic theory**;
 - in other cases, these are **omnibus measures** that **collapse** the use of multiple linguistic constructs into a single quantitative value.”

Egbert, Larsson, & Biber (2020)

WHY DOES IT MATTER?



NEXT UP

Illustrate some of the potential problems of relying on opaque measures that are automatically calculated by corpus analysis software

Case study: Larsson & Kaatari (2020)





DOES IT MATTER?


CASE STUDY: SYNTACTIC COMPLEXITY

SYNTACTIC COMPLEXITY

Syntactic complexity

- the addition of optional structural elements to 'simple' phrases and clauses (e.g., Biber et al., 2022)

More complex

- 
- There is a **strong** need for **further** high-quality research **into the association between the experience of stress across a variety of contexts**

- There is a need for high-quality research

Less complex

THE LARSSON & KAATARI STUDY

Aim:

- investigate grammatical complexity in learner and expert writing from different registers

Step 1:

- Use an automated tool to explore how the measures were patterned across registers and how they were used in learner writing
- But... the program could not provide sufficient information for a detailed linguistic analysis of the results



THE LARSSON & KAATARI STUDY

Aim:

- investigate grammatical complexity in learner and expert writing from different registers

Step 1:

- Use an automated tool to explore how the measures were patterned across registers and how they were used in learner writing
- But... the program could not provide sufficient information for a detailed linguistic analysis of the results



20

Step 1.1:

- Use the online mode of the program
 - allows for sentence-by-sentence tagging
- to try to isolate the different measures and thus decode the numeric scores
- Still several questions remained unanswered
 - in part because interpreting the measures themselves proved challenging...
- Example: Complex nominals per T-unit
 - one of the most important predictor measures in the Larsson & Kaatari study...



COMPLEX NOMINALS PER T-UNIT

- A ratio-based measure
 - Challenging (stay tuned for more on that!)
- Even in isolation, the numerator (complex nominals) and the denominator (T-units) pose problems for the linguistic interpretability of the results



COMPLEX NOMINALS

A measure that covers structures including

- nouns plus adjectives
- possessives
- prepositional phrases
- relative clauses
- participles
- appositives
- nominal clauses (complement clauses controlled by verbs)
- gerunds and infinitives when found in subject position (Lu, 2010: 483)

Confounds A LOT of structurally and syntactically distinct grammatical features! (see Biber et al., 2020)

COMPLEX NOMINALS

- In addition, the exploratory analysis indicated that the measure was **dichotomous**
- = a noun phrase was coded as a “complex nominal” if it had any of the characteristics

A. The **green** book was written in 1953.

CN: 1.0	pre-modification
---------	-------------------------

B. The book **[which is very interesting]** was written in 1953.

CN: 1.0	post-modification
---------	--------------------------

C. The **green** book **[which is very interesting]** was written in 1953.

CN: 1.0	pre and post-modification
---------	----------------------------------

COMPLEX NOMINALS

What are some issues?

- it doesn't distinguish between pre- and post-modification
- nor between single and multiple modification

Also

- The measure is given a label that inaccurately suggests a clear linguistic interpretation
- it is nearly impossible to evaluate the actual linguistic basis of the measure as applied to specific texts

T-UNITS

- Arguably less problematic, but...
- Defined as “one main clause plus any subordinate clause or non-clausal structure that is attached to or embedded in it” (Hunt, 1970: 4)



T-UNITS

Issue:

- T-unit measures **conflate different structural and syntactic characteristics** and are thus very difficult to interpret linguistically
1. The thing that we often forget to think about was that the place where people made these interactions musically was out in the fields.
 2. There is a need for further high-quality research into the association between the experience of stress across a variety of contexts and miscarriage risk.

A single T-unit of the same length:

1. The thing that we often **forget** to **think** about **was** that the place where people **made** these interactions musically **was** out in the fields.

- Spoken interview
- 1 main clause + 4 dependent clauses
- **Extensive clausal elaboration**: a *to*-complement clause, a *that*-complement clause, and two relative clauses

2. There **is** a need for further high-quality research into the association between the experience of stress across a variety of contexts and miscarriage risk.

- Written news article
- 1 main clause + several embedded prepositional phrases
- **Phrasal compression**: multiple phrasal noun modifiers (attributive adjectives, pre-modifying nouns, and post-modifying prepositional phrases)

COMPLEX NOMINALS **PER** T-UNIT

- A ratio-based measure
 - Challenging (stay tuned for more on that!)
- The score is an amalgam of the individual scores from the numerator and the denominator
 - linguistic interpretation requires a separate evaluation of the score for the number of complex nominals and the score for the number of T-units



COMPLEX NOMINALS PER T-UNIT VS. PER CLAUSE

- In expert writing: CN per T-unit and CN per clause correlated at $r = .93$ (Larsson & Kaatari, 2020)

Yet,

Expert writers: a higher average ratio of CN per clause than the learners

Expert writers: minimal differences vis-à-vis the learners for CN per T-unit

- How can that be? And why is that?

A MYSTERY-SOLVING EXPEDITION



- All things being equal, **fewer** dependent clauses in the expert data than in the learner data might possibly explain why the measures were strongly correlated in the expert data, as this would bring scores for clause-based measures closer to those of T-unit-based measures.
 - BUT that was not the case...
- Online mode of the program + manual investigation →

Possible solution:

- the extent to which structures classified as complex nominals were dispersed evenly across clauses

THOUGH NB!



- These steps still did not provide a clear answer to the question of how the language of the learners differed from that of the experts, as the complex nominals measure confounds multiple linguistic structures.
- We carried out complementary manual, computational, and statistical analyses
- Main differences between the experts and the learners:
 - The use of prepositional and adjectival modifiers
 - The experts used a denser style of writing involving more complex noun phrases with pre- and post-modification, in line with previous research on academic writing

TAKE-HOME MESSAGES FROM/FOR LARSSON & KAATARI

Trying to interpret results that stem from automatically calculated measures that are linguistically opaque is a **cumbersome** and, in many cases, even **futile** process

Tools may be very easy to use and give the appearance of carrying out a sophisticated corpus analysis

But! The measures provided are often

- **linguistically uninterpretable** and
- **cannot be evaluated for their linguistic accuracy**



SUGGESTED WAYS FORWARD

A CONCRETE SUGGESTION

Opt for a simpler analysis if need be

...with the primary goal of ensuring an accurate analysis that is **directly interpretable relative to the linguistic research questions of interest**

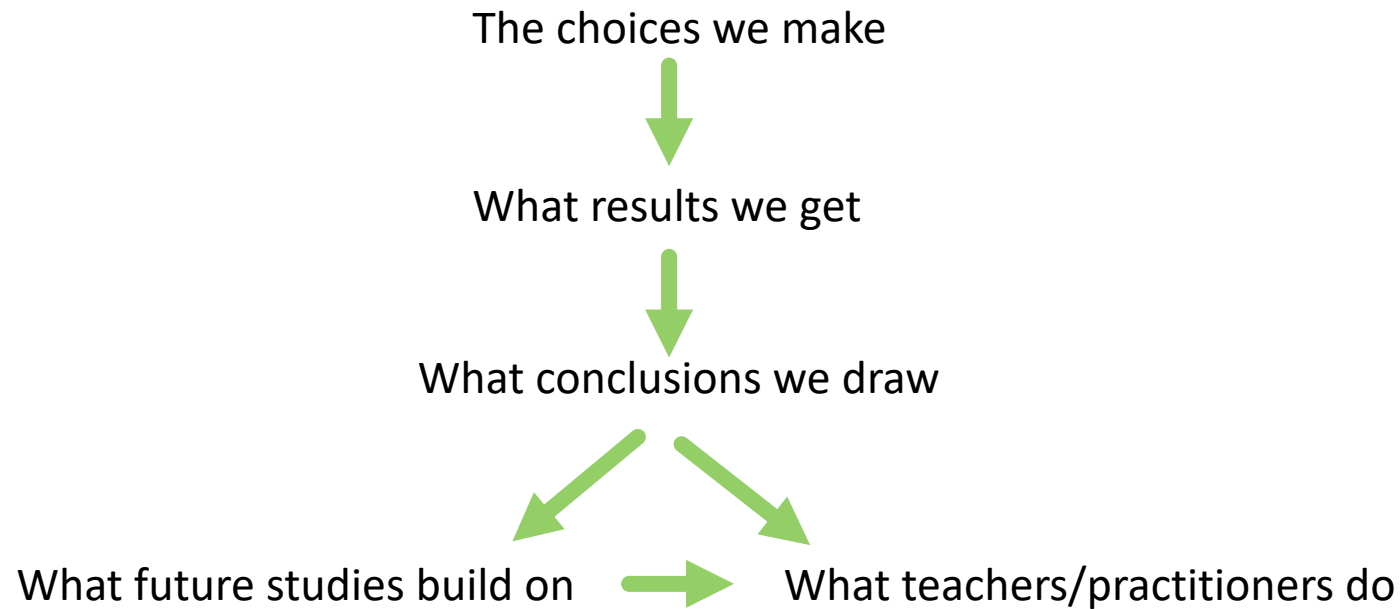


THE IMPORTANCE OF OPERATIONAL DEFINITIONS

- Only with clear operational definitions can we
 - reach higher inter (and intra!) rater agreement (manual investigation)
 - obtain higher accuracy for our measures (computational tools)
 - draw more reliable conclusions from our findings
- The importance of
 - measuring (and reporting!) inter-rater agreement
 - measuring (and reporting!) precision and recall = accuracy

Larsson, Paquot, & Plonsky (2020); Egbert et al. (2020)

WHY DOES IT MATTER?



The choices we make not only as researchers, but also as reviewers, *consumers*, etc.!

CONCLUSION

CONCLUSION

- Opaque measures and methods may leave us unable to draw reliable conclusions from our analysis
- Let's do our best to work toward increased linguistic interpretability, transparency, and rigorousness





THANK YOU!

Tove.Larsson@nau.edu
www.tovelarssoncl.wordpress.com
@tovelarsson1