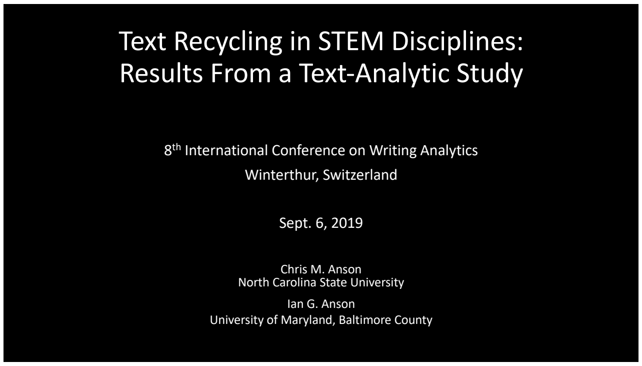


Text Recycling in STEM Disciplines: Results From a Text-Analytic Study

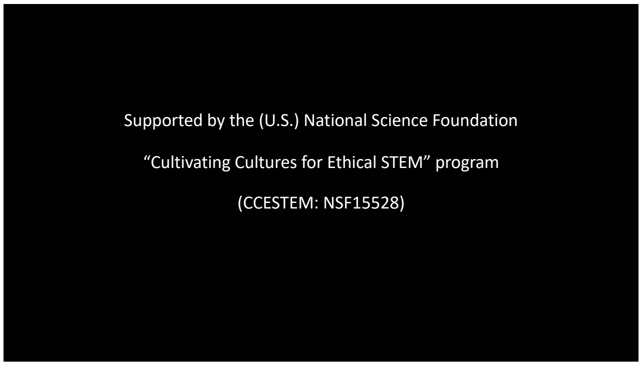
8th International Conference on Writing Analytics
Winterthur, Switzerland

Sept. 6, 2019

Chris M. Anson
North Carolina State University

Ian G. Anson
University of Maryland, Baltimore County

1

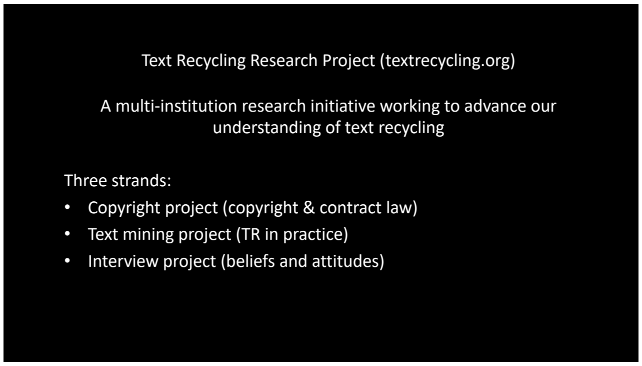


Supported by the (U.S.) National Science Foundation

"Cultivating Cultures for Ethical STEM" program

(CCESTEM: NSF15528)

2

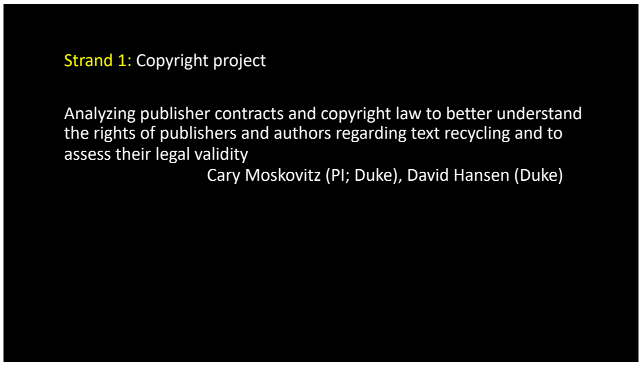


Text Recycling Research Project (textrecycling.org)

A multi-institution research initiative working to advance our understanding of text recycling

Three strands:
- Copyright project (copyright & contract law)
- Text mining project (TR in practice)
- Interview project (beliefs and attitudes)

3

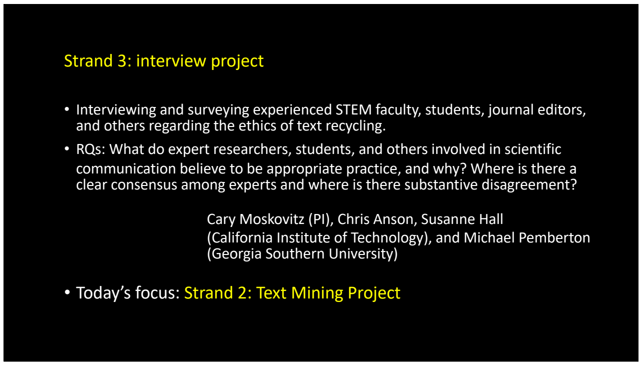


Strand 1: Copyright project

Analyzing publisher contracts and copyright law to better understand the rights of publishers and authors regarding text recycling and to assess their legal validity

Cary Moskovitz (PI; Duke), David Hansen (Duke)

4



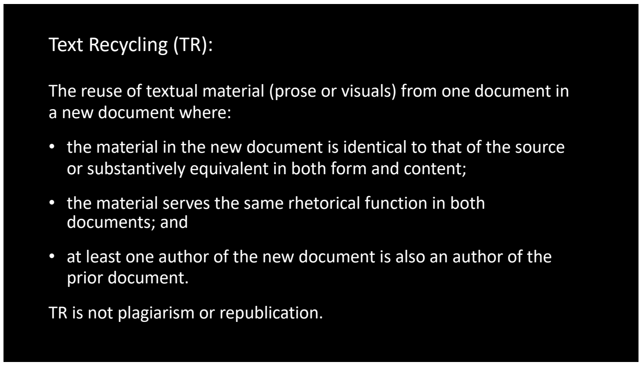
Strand 3: interview project

- Interviewing and surveying experienced STEM faculty, students, journal editors, and others regarding the ethics of text recycling.
- RQs: What do expert researchers, students, and others involved in scientific communication believe to be appropriate practice, and why? Where is there a clear consensus among experts and where is there substantive disagreement?

Cary Moskovitz (PI), Chris Anson, Susanne Hall
(California Institute of Technology), and Michael Pemberton
(Georgia Southern University)

- Today's focus: Strand 2: Text Mining Project

5



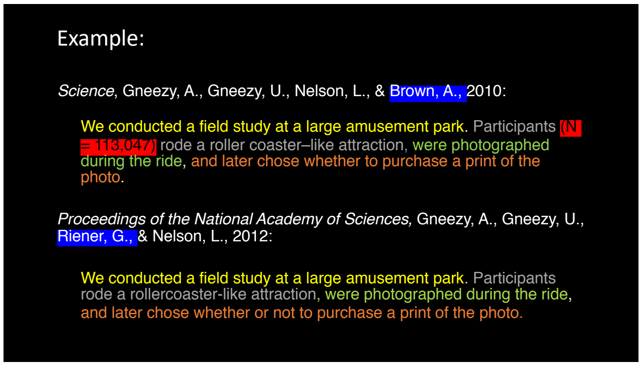
Text Recycling (TR):

The reuse of textual material (prose or visuals) from one document in a new document where:

- the material in the new document is identical to that of the source or substantively equivalent in both form and content;
- the material serves the same rhetorical function in both documents; and
- at least one author of the new document is also an author of the prior document.

TR is not plagiarism or republication.

6



Example:

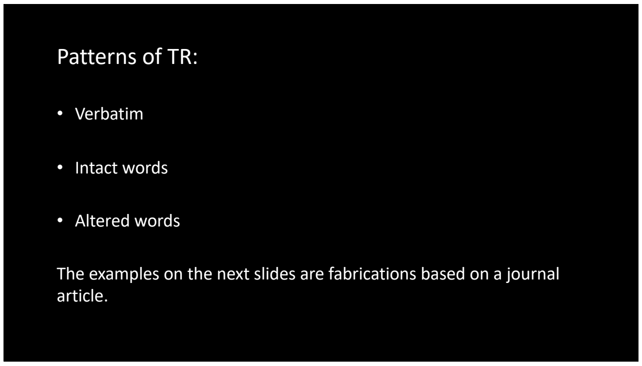
Science, Gneezy, A., Gneezy, U., Nelson, L., & Brown, A., 2010:

> We conducted a field study at a large amusement park. Participants (N = 113,047) rode a roller coaster–like attraction, were photographed during the ride, and later chose whether to purchase a print of the photo.

Proceedings of the National Academy of Sciences, Gneezy, A., Gneezy, U., Riener, G., & Nelson, L., 2012:

> We conducted a field study at a large amusement park. Participants rode a rollercoaster-like attraction, were photographed during the ride, and later chose whether or not to purchase a print of the photo.

7

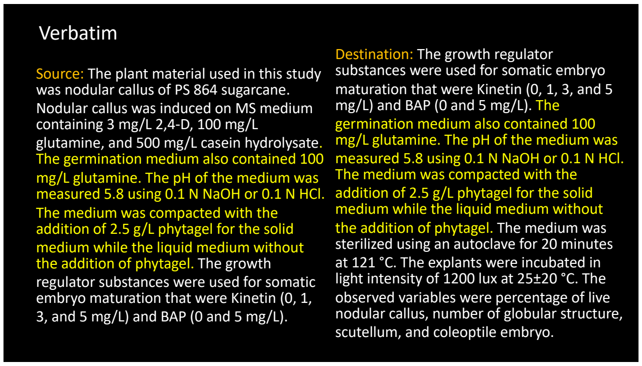


Patterns of TR:

- Verbatim
- Intact words
- Altered words

The examples on the next slides are fabrications based on a journal article.

8

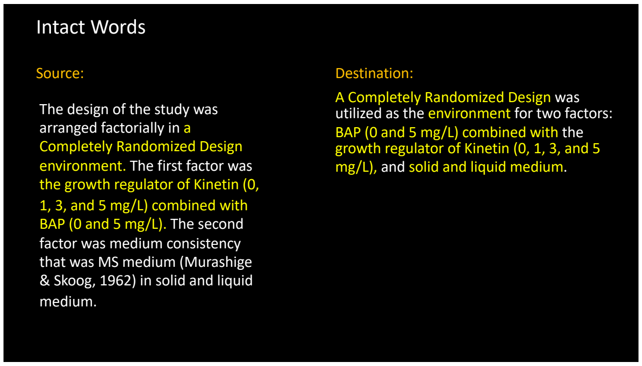


Verbatim

Source: The plant material used in this study was nodular callus of PS 864 sugarcane. Nodular callus was induced on MS medium containing 3 mg/L 2,4-D, 100 mg/L glutamine, and 500 mg/L casein hydrolysate. The germination medium also contained 100 mg/L glutamine. The pH of the medium was measured 5.8 using 0.1 N NaOH or 0.1 N HCl. The medium was compacted with the addition of 2.5 g/L phytagel for the solid medium while the liquid medium without the addition of phytagel. The growth regulator substances were used for somatic embryo maturation that were Kinetin (0, 1, 3, and 5 mg/L) and BAP (0 and 5 mg/L).

Destination: The growth regulator substances were used for somatic embryo maturation that were Kinetin (0, 1, 3, and 5 mg/L) and BAP (0 and 5 mg/L). The germination medium also contained 100 mg/L glutamine. The pH of the medium was measured 5.8 using 0.1 N NaOH or 0.1 N HCl. The medium was compacted with the addition of 2.5 g/L phytagel for the solid medium while the liquid medium without the addition of phytagel. The medium was sterilized using an autoclave for 20 minutes at 121 °C. The explants were incubated in light intensity of 1200 lux at 25±20 °C. The observed variables were percentage of live nodular callus, number of globular structure, scutellum, and coleoptile embryo.

9



Intact Words

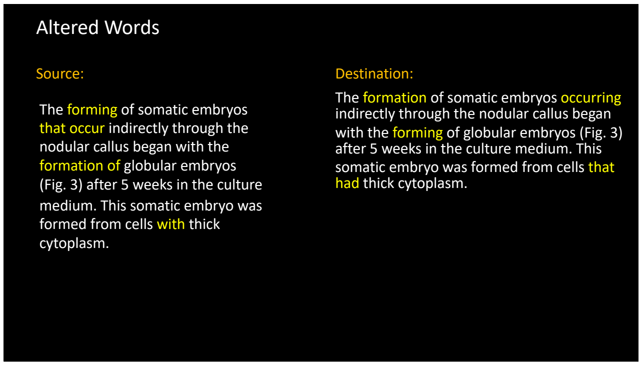
Source:

The design of the study was arranged factorially in a Completely Randomized Design environment. The first factor was the growth regulator of Kinetin (0, 1, 3, and 5 mg/L) combined with BAP (0 and 5 mg/L). The second factor was medium consistency that was MS medium (Murashige & Skoog, 1962) in solid and liquid medium.

Destination:

A Completely Randomized Design was utilized as the environment for two factors: BAP (0 and 5 mg/L) combined with the growth regulator of Kinetin (0, 1, 3, and 5 mg/L), and solid and liquid medium.

10



Altered Words

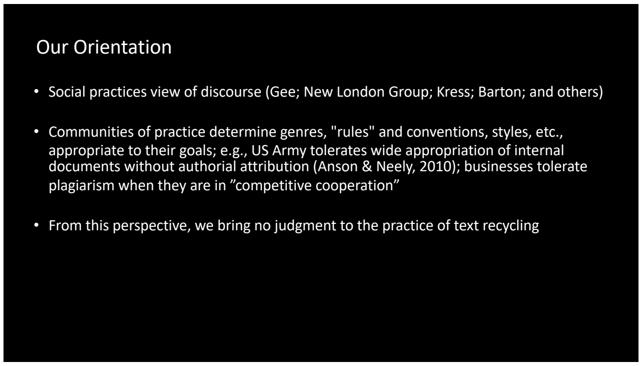
Source:

The forming of somatic embryos that occur indirectly through the nodular callus began with the formation of globular embryos (Fig. 3) after 5 weeks in the culture medium. This somatic embryo was formed from cells with thick cytoplasm.

Destination:

The formation of somatic embryos occurring indirectly through the nodular callus began with the forming of globular embryos (Fig. 3) after 5 weeks in the culture medium. This somatic embryo was formed from cells that had thick cytoplasm.

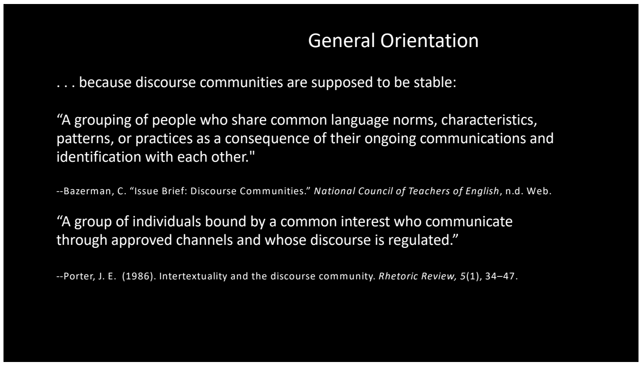
11



Our Orientation

- Social practices view of discourse (Gee; New London Group; Kress; Barton; and others)
- Communities of practice determine genres, "rules" and conventions, styles, etc., appropriate to their goals; e.g., US Army tolerates wide appropriation of internal documents without authorial attribution (Anson & Neely, 2010); businesses tolerate plagiarism when they are in ”competitive cooperation”
- From this perspective, we bring no judgment to the practice of text recycling

12



General Orientation

. . . because discourse communities are supposed to be stable:

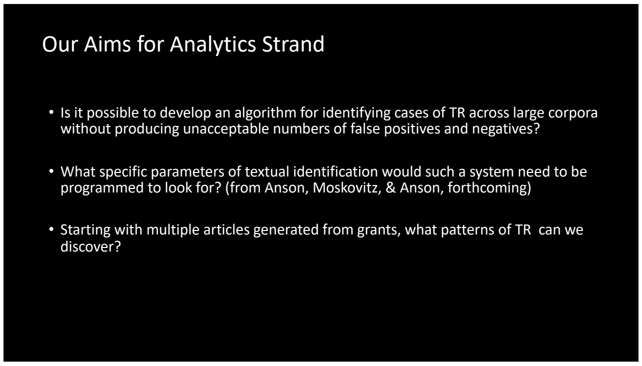
"A grouping of people who share common language norms, characteristics, patterns, or practices as a consequence of their ongoing communications and identification with each other."

--Bazerman, C. "Issue Brief: Discourse Communities." *National Council of Teachers of English*, n.d. Web.

"A group of individuals bound by a common interest who communicate through approved channels and whose discourse is regulated."

--Porter, J. E. (1986). Intertextuality and the discourse community. *Rhetoric Review, 5*(1), 34–47.

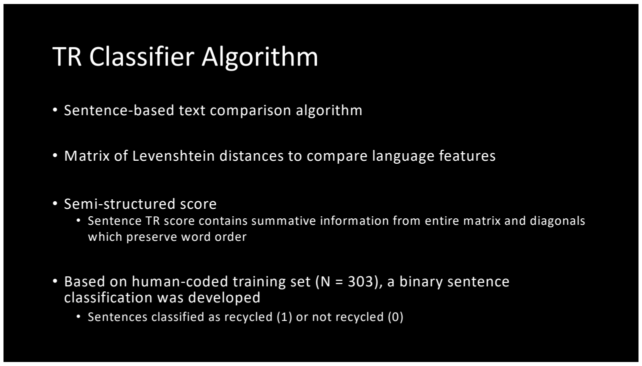
13



Our Aims for Analytics Strand

- Is it possible to develop an algorithm for identifying cases of TR across large corpora without producing unacceptable numbers of false positives and negatives?
- What specific parameters of textual identification would such a system need to be programmed to look for? (from Anson, Moskovitz, & Anson, forthcoming)
- Starting with multiple articles generated from grants, what patterns of TR can we discover?

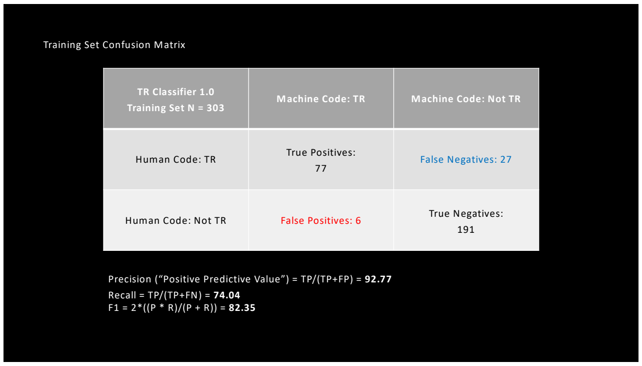
14



TR Classifier Algorithm

- Sentence-based text comparison algorithm
- Matrix of Levenshtein distances to compare language features
- Semi-structured score
  - Sentence TR score contains summative information from entire matrix and diagonals which preserve word order
- Based on human-coded training set (N = 303), a binary sentence classification was developed
  - Sentences classified as recycled (1) or not recycled (0)

15



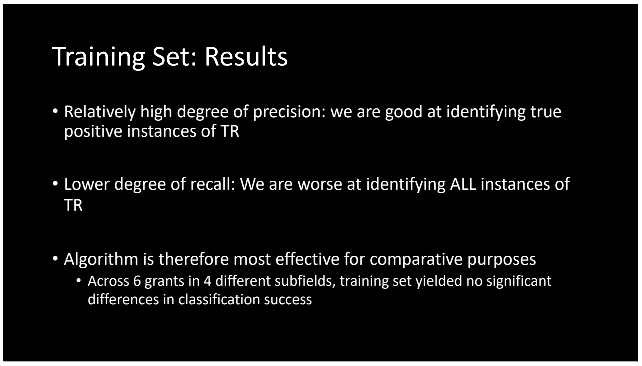
Training Set Confusion Matrix

TR Classifier 1.0 Training Set N = 303	Machine Code: TR	Machine Code: Not TR
Human Code: TR	True Positives: 77	False Negatives: 27
Human Code: Not TR	False Positives: 6	True Negatives: 191

Precision (“Positive Predictive Value”) = TP/(TP+FP) = **92.77**

Recall = TP/(TP+FN) = **74.04**
F1 = 2*((P * R)/(P + R)) = **82.35**

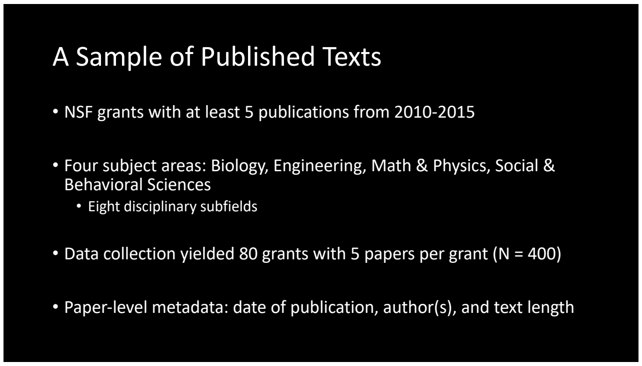
16



Training Set: Results

- Relatively high degree of precision: we are good at identifying true positive instances of TR
- Lower degree of recall: We are worse at identifying ALL instances of TR
- Algorithm is therefore most effective for comparative purposes
  - Across 6 grants in 4 different subfields, training set yielded no significant differences in classification success

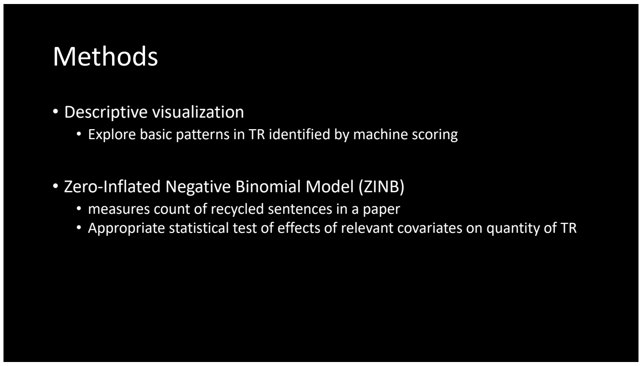
17



A Sample of Published Texts

- NSF grants with at least 5 publications from 2010-2015
- Four subject areas: Biology, Engineering, Math & Physics, Social & Behavioral Sciences
  - Eight disciplinary subfields
- Data collection yielded 80 grants with 5 papers per grant (N = 400)
- Paper-level metadata: date of publication, author(s), and text length

18



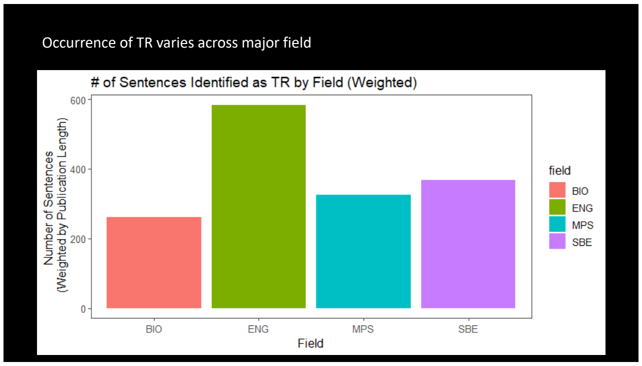
Methods

- Descriptive visualization
  - Explore basic patterns in TR identified by machine scoring
- Zero-Inflated Negative Binomial Model (ZINB)
  - measures count of recycled sentences in a paper
  - Appropriate statistical test of effects of relevant covariates on quantity of TR

19

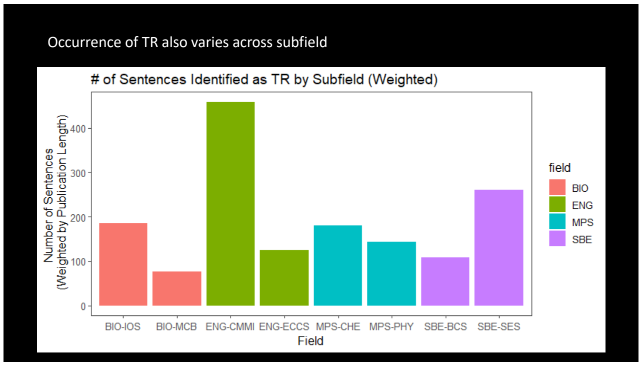
Results

20



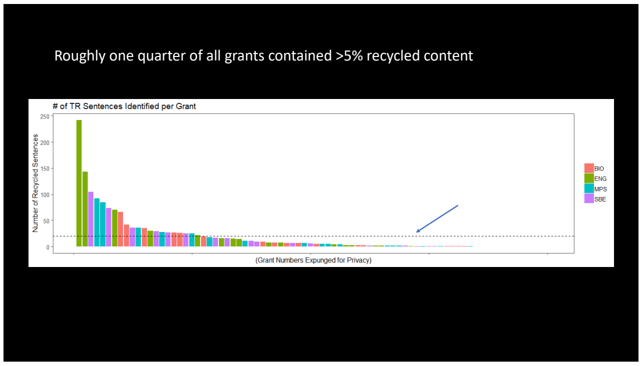
Occurrence of TR varies across major field

21



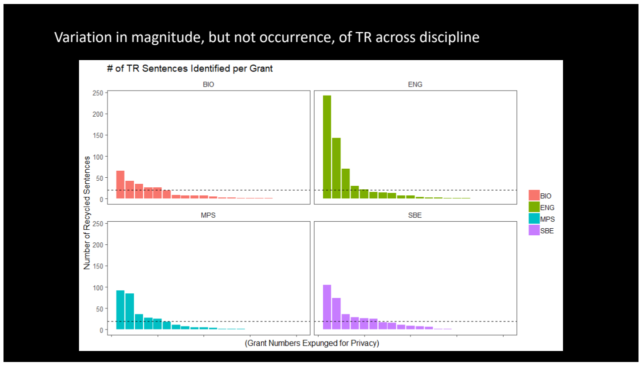
Occurrence of TR also varies across subfield

22



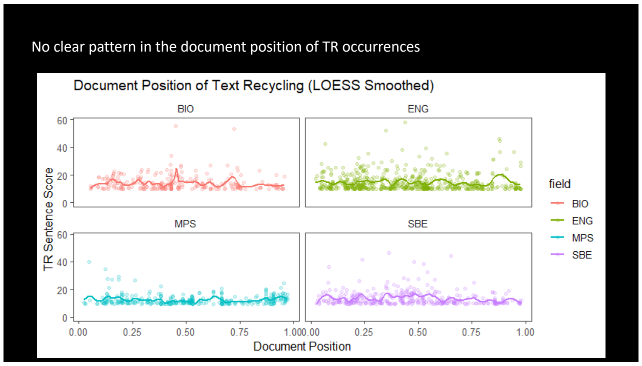
Roughly one quarter of all grants contained >5% recycled content

23



Variation in magnitude, but not occurrence, of TR across discipline

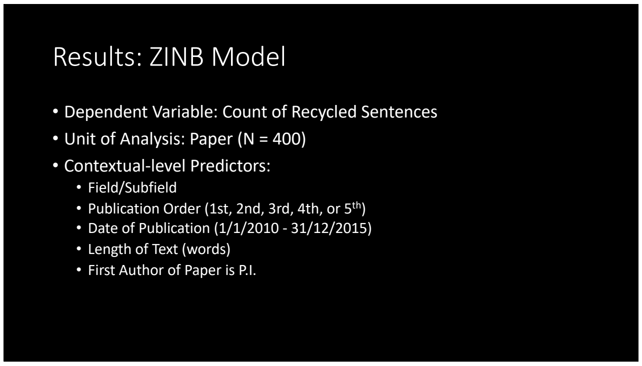
24



No clear pattern in the document position of TR occurrences

Document Position of Text Recycling (LOESS Smoothed)

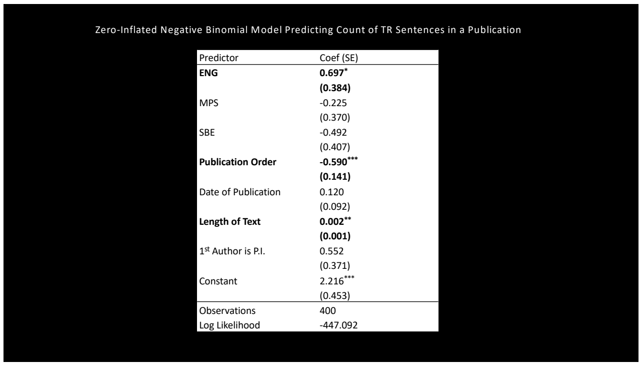
25



Results: ZINB Model

- Dependent Variable: Count of Recycled Sentences
- Unit of Analysis: Paper (N = 400)
- Contextual-level Predictors:
  - Field/Subfield
  - Publication Order (1st, 2nd, 3rd, 4th, or 5th)
  - Date of Publication (1/1/2010 - 31/12/2015)
  - Length of Text (words)
  - First Author of Paper is P.I.

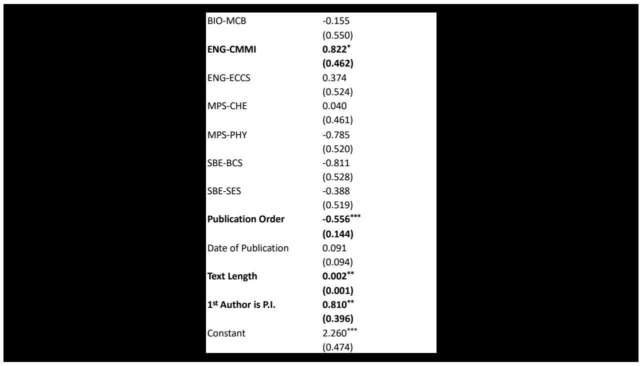
26



Zero-Inflated Negative Binomial Model Predicting Count of TR Sentences in a Publication

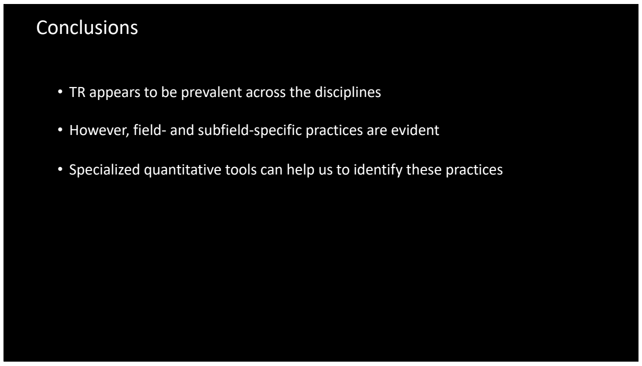
Predictor	Coef (SE)
ENG	**0.697***
	(0.384)
MPS	-0.225
	(0.370)
SBE	-0.492
	(0.407)
Publication Order	**-0.590*****
	(0.141)
Date of Publication	0.120
	(0.092)
Length of Text	**0.002****
	(0.001)
1st Author is P.I.	0.552
	(0.371)
Constant	2.216***
	(0.453)
Observations	400
Log Likelihood	-447.092

27



BIO-MCB	-0.155
	(0.550)
ENG-CMMI	**0.822***
	(0.462)
ENG-ECCS	0.374
	(0.524)
MPS-CHE	0.040
	(0.461)
MPS-PHY	-0.785
	(0.520)
SBE-BCS	-0.811
	(0.528)
SBE-SES	-0.388
	(0.519)
Publication Order	**-0.556*****
	(0.144)
Date of Publication	0.091
	(0.094)
Text Length	**0.002****
	(0.001)
1st Author is P.I.	**0.810****
	(0.396)
Constant	2.260***
	(0.474)

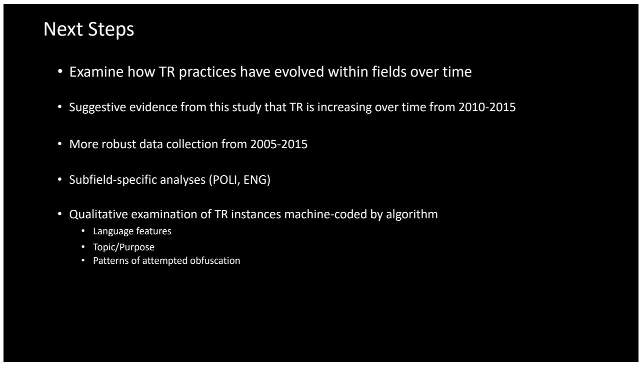
28



Conclusions

- TR appears to be prevalent across the disciplines
- However, field- and subfield-specific practices are evident
- Specialized quantitative tools can help us to identify these practices

29



Next Steps

- Examine how TR practices have evolved within fields over time
- Suggestive evidence from this study that TR is increasing over time from 2010-2015
- More robust data collection from 2005-2015
- Subfield-specific analyses (POLI, ENG)
- Qualitative examination of TR instances machine-coded by algorithm
  - Language features
  - Topic/Purpose
  - Patterns of attempted obfuscation

30

Questions and Discussion

31

Merci vilmal!

Chris Anson
chris_anson@ncsu.edu
www.ansonica.net

Ian Anson
iganson@umbc.edu
www.iananson.com

32