Teaching Scientific Writing at Scale: Characterizing Student Writing in Undergraduate Biology

A. Daniel Johnson[1], PhD, Sabrina D. Setaro[1], PhD, Jerid C. Francom[2], PhD
Depts. of Biology[1], Linguistics[2], Wake Forest University, Winston-Salem, NC

Overview

Needs, Challenges

- Scientific writing builds students' thinking & communication skills.
- Integrating writing into large BIO101 classes is challenging.
- How can students' writing be evaluated longitudinally in large courses?

Guiding Questions

- Can machine-scorable text features be proxy metrics for students' development as writers?
- What metrics are informative?
- What do they tell us?
- Can they summarize, illuminate cohort-level changes over a course sequence?

Approach

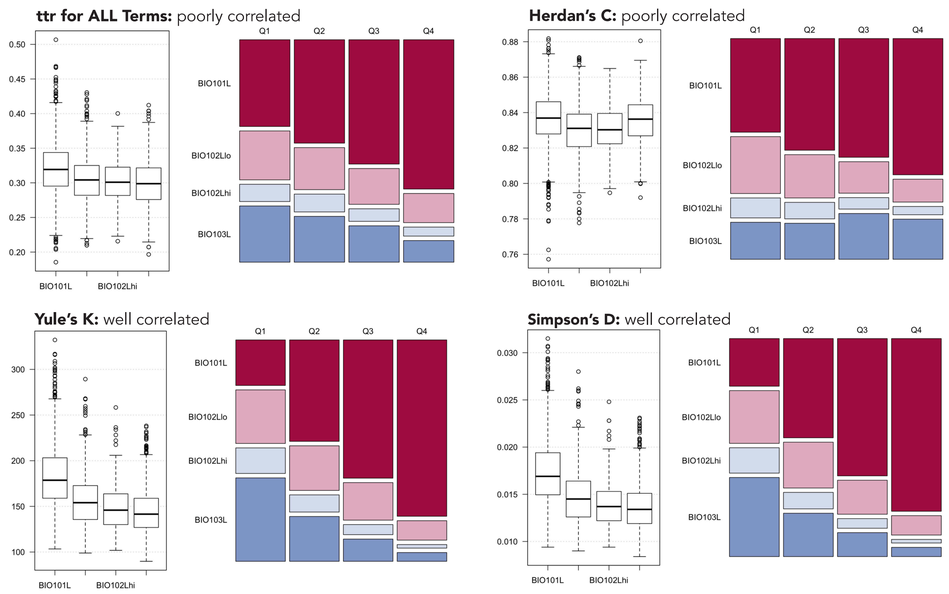
- >4400 student lab reports split into *4 writer experience levels*.
- Reports *bins-scored* by GTAs using *~20 fixed criteria* as:
  - Acceptable
  - Needs minor improvements
  - Needs major improvements
  - Unacceptable/Flawed
- Features scoring:

Lexical Range, Word Choice	Readability		
# word types, tokens; TTRs • VR, TL • ttrFull • ttrStop • ttrGenl • ttrAcad • ttrTech • ttrDisc • Guiraud's R • Herdan's C • Dugast's U Word repetition • Yule's K • Herdan's V_m • Simpson's D	• Automated Readability Index (ARI) • New ARI Bormuth Indices • Mean Cloze • Grade Placement Coleman-Liau Indices • Cloze • Grade Placement RIX • Index score • Grade level	• Fog - Powers, Sumner, Kearl • New Fog Count • SMOG Index • Gunning Fog Index • Flesch Reading Ease • Flesch-Kincaid • Flesch (Powers, Sumner, Kearl) • Farr, Jenkins, Paterson (FJP) • New FJP • FJP - Powers, Sumner, Kearl • Lensear Write	• % Dale-Chall unfamiliar words • % Fog Hard Words • # long words (6+ chars) • # passive constructions • % SMOG "hard" words • # wordy items • Mean sentence length • Mean word syllables

  - *Vocabulary range/richness* = # unique words, type token ratios & variants
  - *Word choices* classified by fixed vocabularies.
  - *Readability indices* = sentence complexity.
- *Proportional odds ordinal logistic regression (POLR)*: can proxy metrics predict human-assigned scores?

Key Findings

Several machine-scored metrics correlated well with students' growing experience as writers

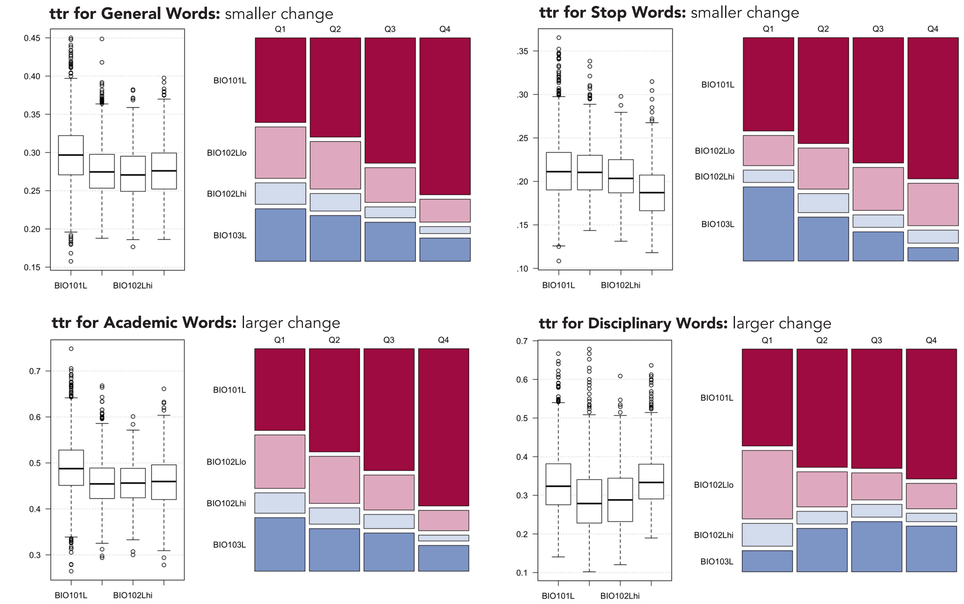


		Length Adjusted ttr Variants			Measures of Repetition		
Course	Simple TTR	Guiraud's R	Herdan's C	Dugast's Index (U)	Yule's K	Herdan's V_m	Simpson's D
Bio101	0.320±0.038	10.4± 1.02	0.837± 0.014	18.6± 1.42	182.6± 33.4	0.123± 0.013	0.017± 0.003
BIO103lo	0.305±0.035 [A]	10.1± 1.02 [A]	0.830± 0.014 [A]	18.0± 1.38 [A]	157.4± 28.7 [A]	0.112± 0.012[A]	0.015± 0.003[A]
Bio103hi	0.303±0.032	10.3± 0.92	0.831± 0.013	18.1± 1.25	149.0± 25.2 [B]	0.109± 0.011[B]	0.014± 0.003[B]
Bio102	0.299±0.043 [C]	11.3± 1.02 [C]	0.835± 0.013 [C]	19.2± 1.31 [C]	145.0± 25.6	0.109± 0.011	0.014± 0.003
% change	-6.6% ***	+8.7% ***	-0.02%***	+3.2% ***	-20.6% ***	11.4%***	-17.6% ***
ϕ_C	0.24	0.37	0.19	0.27	0.49	0.46	0.49

Overall lexical richness (simple type-token ratio, Herdan's C, Dugast's U) did not change with experience.

Word repetition (Yule's K, Simpson's D, Herdan's V_m) declined 11.4-20.6% ($p<0.001$).

Lexical range & use of formal terms increased as students gained writing experience



		Non-Specialized Vocabulary				Specialized Vocabulary					
	Full			Stop		Genl		Acad		Tech		Disc	
Course	VR	TL	ttr	VR	ttr	VR	ttr	VR	ttr	VR	ttr	VR	ttr
Bio101 (n=2471)	390± 75	1248± 336	0.320± 0.038	116± 20	0.213± 0.034	208± 42	0.297± 0.039	215± 47	0.490± 0.061	21± 5.9	0.498± 0.13	33± 8.2	0.332± 0.079
BIO102Lo (n=831)	386± 72 [A]	1288± 314 [A]	0.305± 0.035 [A]	120± 19 [A]	0.212± 0.031	206± 40	0.278± 0.036 [A]	214± 47	0.457± 0.054 [A]	21± 6.6	0.507± 0.146	27± 6.8 [A]	0.291± 0.09 [A]
Bio102Hi (n=302)	398± 66 [B]	1335± 306 [B]	0.303± 0.032	118± 18	0.207± 0.029	202± 34	0.273± 0.035	228± 45 [B]	0.455± 0.049	23± 7 [B]	0.499± 0.137	29± 6.7 [B]	0.293± 0.084
Bio103 (n=853)	488± 90 [C]	1673± 446 [C]	0.299± 0.043 [C]	130± 20 [C]	0.188± 0.03 [C]	243± 47 [C]	0.276± 0.036	267± 55 [C]	0.459± 0.057	29± 7 [C]	0.481± 0.095	44± 11 [C]	0.340± 0.072 [C]
% change	25.1% ***	34.1% ***	-6.6% ***	12.1% ***	-11.7% ***	17.8% ***	-7.1% ***	24.2% ***	-6.3% ***	38.1% ***	-3.4% (p= 0.04)	33.3% ***	+2.4% ***
Relative assoc. ϕ_C	0.42	0.39	0.24	0.26	0.29	0.31	0.27	0.36	0.28	0.39	0.043 (NS)	0.51	0.28

Type counts (VR), token counts (TL), and type-token ratios (ttr) overall and for sub-vocabularies are reported as mean±st.dev. Percent change is the difference between the first course (BIO101) and final course (BIO102L); overall differences within vocabulary groups are listed with % change (Kruskal-Wallis rank sum test, *** = p<0.001.) Within-group differences were evaluated using Dunn's test ([A]=p<0.01 for BIO101L vs BIO102Lo, [B]=p<0.01 for BIO102Lo vs BIO102Lhi, [C]=p<0.01 for BIO102Lhi vs. BIO103.) Relative associations within each group were estimated using Cramér's phi (ϕ_C); only groups with $\phi_C > 0.1$ were considered further. Full = values for full text, Stop = only terms in stopwords sub-vocabulary, Genl = general sub-vocabulary, Acad = academic sub-vocabulary, Tech = technical sub-vocabulary, Disc = discipline-specific sub-vocabulary.

Total # unique words used rose 25.1% ($p<0.001$).

Use of academic & specialized terms grew faster (24.2-38.1%) than general terms (12.1%-17.8%), reflecting a move to more "formal" word choices.

Readability indices stressing poly-syllaby correlated better with writing experience

New Fog Count: a poorly correlated metric

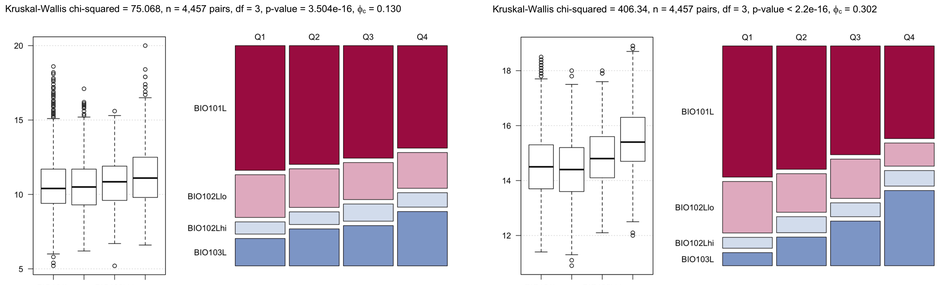
	BIO101L	BIO102L.low	BIO102L.high	BIO103L
Mean	10.599	10.561	10.820	11.235
Median	10.400	10.500	10.850	11.100
S.D.	1.833	1.859	1.784	1.960
IQR	2.300	2.400	2.300	2.700

Kruskal-Wallis chi-squared = 75.068, n = 4,457 pairs, df = 3, p-value = 3.504e-16, ϕ_C = 0.130

SMOG Index: a better correlated metric

	BIO101L	BIO102L.low	BIO102L.high	BIO103L
Mean	14.557	14.390	14.895	15.484
Median	14.500	14.400	14.800	15.400
S.D.	1.189	1.114	1.210	1.187
IQR	1.600	1.600	1.500	1.600

Kruskal-Wallis chi-squared = 406.34, n = 4,457 pairs, df = 3, p-value < 2.2e-16, ϕ_C = 0.302



14/32 indices had relative association (ϕC) > 0.2 over 3-course series ($p<0.001$).

Indices stressing wordy items, frequency of long or polysyllabic words rose 10.3-41.9% overall, & were more likely to be correlated with gains in student experience.

Bin Score ~ [Simpson's D + Yule's K] (word repetition)
+ [ttrGenl + ttrAcad] (word choices)
+ [# word types + # tokens] (lexical range, text length)
+ [% SMOG hard words + mean syllables/ word + # 6-character words] (word complexity)
+ [Bormuth Grade Placement + Flesch-Kincaid + Gunning Fog + SMOG] (overall readability)

Proxy metrics were poor predictors of individual grades. Fit for single- & multi-factor POLR models was low, with 59% average predictive error on the best fit model (above; Nagelkerke pseudo-R^2 = 0.187.)

Broader Impacts

- Tracking student development as scientific writers by "close reading" is impractical in high enrollment STEM courses.
- These proxy metrics surface changes in students' writing:
  - Longitudinally over a curriculum sequence, and
  - At a cohort-level scale

Text Feature Groups	Proxy Measures
1. Vocabulary range, text length	• Total # word types (VR) • Total # word tokens (TL) • ttrFull
2. Word repetition	• Simpson's D • Yule's K
3. Word usage choice	• Non-specialized words (ttrGenl, ttrStop) • Specialized words (# wordy items, ttrAcad, ttrDisc)
4. Word complexity	• % SMOG/Fog hard words (3+ syllables) • # 6-character words • Avg. # syllables/word
5. Readability	• Lensear Write, SMOG Index (avg. # 1-, 3+-syllable words/sentence) • Flesch-Kincaid, Flesch-PSK (avg. # syllables/word, avg. # words/sentence) • Gunning Fog, Fog-PSK (% 3+ syllable words, avg. # words/sentence) • ARI, New ARI (avg. # chars./word, avg. words/sentence)

- Proxy metrics analysis of full documents (vs. text samples) is:
  - Scalable
  - Less subject to interpretation
  - Harder to "game"
  - Able to triangulate on writing features of interest/value that instructors want students to develop over time.

Resources to Share

- Corpus + metadata for 4400 student reports
- Scientific Writing Resource Guide (open-source)
- Bins-based scoring rubrics + instructor training materials
- R Shiny form for collecting well-structured student reports
- Structured vocabularies, code-books, & R scripts for analyses.

Find us online at:
https://github.com/adanieljohson/stemwritingproject
OR
https://qubeshub.org/community/groups/stemwritingproject

Disclaimer: This material is based on work supported by the National Science Foundation under DUE Grant #**1712423**. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation.