Schwa Density Paper

Kyle Townsend

Claude (analysis collaborator)

2026-04-16

Abstract

We test whether schwa density, the proportion of vowel phones in a text that are unstressed schwa (CMUdict `AH0`), can serve as a phonologically motivated single-feature register classifier in English text. A pre-registered confirmatory plan locking three hypothesis tests (minimum effect, non-inferiority versus Flesch–Kincaid grade level, and correlation replication) was applied to two pre-registered corpora and two additional sensitivity corpora: NLTK multi-source ($N=164$), the Standardized Project Gutenberg Corpus ($N=2{,}767$), the Brown corpus ($N=313$ qualifying), and the Open American National Corpus ($N=4{,}375$). Both pre-registered confirmatory tests passed: schwa density discriminated registers above the crud floor and matched Flesch–Kincaid within the pre-specified non-inferiority margin. Joint partial-η^2 controls for syllables per word, mean word length, and Latinate ratio show that schwa retains 46–53% of its register-discrimination on the two pre-registered corpora even after absorbing surface lexical signal, supporting the stronger phonological content claim. A function-word ablation (masking the 198 NLTK English stopwords before computing schwa density) preserves or amplifies register discrimination on all four corpora (η^2 retention 0.93–1.27), ruling out function-word frequency as a confound. The ablation result formalises the paper's principal finding: schwa density operates as a *Primary Stylistic Feature* (NLTK, SPGC, Brown; content-word phonological variation, unmasked by stopword removal) and as a *Secondary Modality Feature* (OANC; partially dependent on function-word frequency, slightly attenuated by masking). The measure is most valuable for within-prose stylistic discrimination and degrades to a syllables-per-word proxy when register variation reflects cross-modality contrasts (speech versus technical writing).

Introduction

Stylometric register classification draws from a well-developed toolkit: Biber's multi-dimensional analysis (Biber, 1988), Flesch–Kincaid (FK) grade level (Kincaid et al., 1975), the SMOG index (McLaughlin, 1969), and more recent supervised approaches grounded in lexical and syntactic features (Grieve, Clarke, & Chiang, 2023). What this toolkit lacks is a phoneme-level single-feature measure that is robust to preprocessing variation and conceptually transparent.

We propose schwa density, defined as the proportion of vowel phones in a text that are unstressed schwa (`AH0` in the CMU Pronouncing Dictionary), as such a measure. The measure has two attractive properties: it depends only on a phonemic dictionary lookup (no sentence boundary detection required), and it is grounded in a well-studied phonological process (English vowel reduction) rather than orthographic surface statistics.

This paper reports a four-corpus pre-registered confirmatory test of the claim that schwa density is competitive with Flesch–Kincaid grade level as a single-feature register classifier. The pre-registration was locked before any confirmatory data look. We honour those locked tests in the results section and use the writeup to position the finding within its actual scope: schwa is most useful for within-prose stylistic discrimination and degrades to a syllables-per-word proxy when register variation is driven by cross-modality contrasts.

Method

Schwa density operationalisation

For each text, we compute the phoneme sequence by mapping each alphabetic word to its first CMUdict pronunciation (CMU, 1998).[1] We extract the vowel phones, identifying each by base (one of the 14 ARPAbet vowels) and primary stress digit. The primary measure ($\mathrm{schwa}_{\mathrm{v1}}$) is the proportion of vowel phones equal to `AH0`. Three sensitivity variants are computed and reported but not selectable post hoc: $\mathrm{v2}$ = (`AH0`+`IH0`)/total, $\mathrm{v3}$ = all unstressed vowels (any stress digit 0)/total, $\mathrm{v4}$ = any stress level `AH`/total.

Comparison features

For each text we additionally compute mean syllables per word (CMUdict syllable counts with a heuristic fallback for OOV words), mean word length in characters, mean sentence length in words, type–token ratio, Latinate-ending ratio (count of words ending in any of *tion, ity, ance, ence, ous, ment, ive, al, ary, ory, ism, ist*, divided by word count), conditional vowel transition entropy $H(V_n \mid V_{n-1})$, marginal vowel entropy $H(V)$, and Flesch–Kincaid grade level $\mathrm{FK} = 0.39\,\overline{\ell_s} + 11.8\,\overline{\ell_w} - 15.59$, where $\overline{\ell_s}$ is mean sentence length in words and $\overline{\ell_w}$ is mean syllables per word.

Corpora

We test on four corpora spanning four sourcing conventions:

Brown corpus ($N{=}500$; 313 qualifying after the pre-registered $N \geq 30$ bucket-exclusion rule). Used as exploratory reference. Bundled with NLTK (Bird et al., 2009). We report results on both the locked 6-bucket grouping (the categories that meet $N \geq 30$ on this collection) and on a standard 5-bucket grouping (press, general, learned, fiction, religion) for comparison with prior reported numbers.

NLTK multi-source ($N{=}164$; 135 qualifying). Pre-registered confirmatory corpus, assembled per the locked stratification: literary fiction (Gutenberg), drama (Shakespeare), oratorical (inaugural and state-of-the-union addresses), news (Reuters and ABC), reviews (movie reviews), and web-informal (web text and chat). Texts shorter than 1,000 alphabetic tokens were either rejected or batched to clear the floor.

Standardized Project Gutenberg Corpus (SPGC) (Gerlach & Font-Clos, 2020) ($N{=}2{,}767$; all 12 buckets qualify). The largest pre-registered confirmatory corpus. SPGC ships pre-tokenised as one word per line with punctuation and numbers stripped. We patched the analyser to bypass NLTK tokenisation in this case, with the constant $\overline{\ell_s} = 20$ heuristic for FK (sentence boundaries are unrecoverable from the token stream). Subjects in the SPGC metadata file are LCSH headings rather than LCC classification codes (Library of Congress(n.d.)); we constructed register buckets by regex on LCSH first-segments (priority: children, drama, poetry, religion, philosophy, science, history, biography,

travel, essays, letters, fiction; multi-match resolved by priority with `fiction` as catch-all). This is documented as a deviation from the original handoff.

Open American National Corpus (OANC-GrAF) (Ide & Suderman, 2007) ($N = 4{,}375$; 6 qualifying buckets). Added as non-pre-registered sensitivity corpus. Register taxonomy taken from OANC's directory hierarchy: face-to-face conversation, telephone conversation, journal (Slate magazine + Verbatim), letters, non-fiction, technical (911 report, biomed, government documents, PLOS), and travel guides.

Pre-registered tests

Three tests were locked before any data look on the confirmatory corpora. Bootstrap CIs use 1,000 within-group resamples with `random_state=42`.

T1 (minimum effect).

$H_0: \eta^2(\text{schwa}, \text{register}) \leq 0.04$. Reject if the lower bound of the 95% bootstrap CI for η^2 exceeds 0.04. The 0.04 floor follows Lakens (2022) on minimum-effect-of-interest tests for text-derived measures, which non-trivially correlate at baseline.

T2 (non-inferiority versus FK).

$H_0: \eta^2(\text{schwa}) - \eta^2(\text{FK}) \leq -0.05$. Reject if the lower bound of the 90% bootstrap CI for the difference exceeds -0.05. The 0.05 margin is a cost–benefit-style SESOI: at $\eta^2 \approx 0.5$ it is approximately a 9% relative gap, the smallest gap that would meaningfully recommend FK over schwa given schwa's conceptual simplicity.

T3 (correlation replication).

$H_0: |r(\text{schwa}, \text{cond } H)| \leq 0.36$. Reject if observed $|r| > 0.36$ and one-sided $p < 0.05$. The 0.36 threshold is the small-telescopes value (Simonsohn, 2015): the smallest r that would have reached significance in the original $N = 30$ pilot study.

We frame T3 in the discussion as a pipeline-consistency check rather than substantive evidence about vowel transition structure. Marginal and conditional vowel entropies correlate at

$r \approx 0.99$ in prior corpora, and Shannon entropy mechanically drops as one vowel category (schwa) comes to dominate the distribution. The result of T3 confirms that the phonemic pipeline produces the expected mathematical relationship; it does not constitute independent evidence that vowel transition predictability discriminates registers.

Results

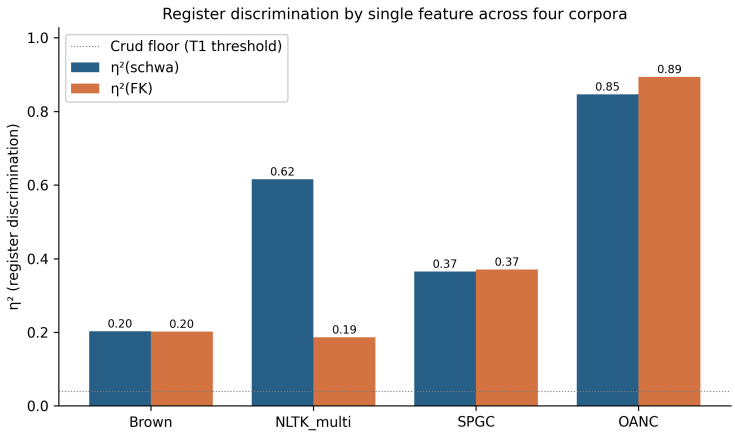
Locked confirmatory tests

Table 1 reports T1–T3 outcomes on all four corpora. Both pre-registered confirmatory corpora (NLTK multi-source and SPGC) cleared all three locked tests. The exploratory reference (Brown) and the sensitivity corpus (OANC) are reported in the same format for context.

Pre-registered tests across four corpora. T1 column reports $\eta^2(\text{schwa})$ with 95% bootstrap CI. T2 column reports $\eta^2(\text{schwa}) - \eta^2(\text{FK})$ with 90% CI. T3 column reports $|r(\text{schwa}, \text{cond } H)|$ (one-sided $p < 10^{-40}$ on all four corpora). Brown is exploratory; NLTK and SPGC are pre-registered confirmatory corpora; OANC is non-pre-registered sensitivity. Brown's T2 failure on the 6-bucket grouping (locked $N \geq 30$ rule) becomes T2 pass on the 5-bucket Francis–Kučera grouping ($\eta^2_{\text{schwa}} = 0.518$, $\eta^2_{\text{FK}} = 0.534$, gap -0.016).

Corpus	N	T1 obs	T2 obs	$\lvert r \rvert$	Outcome
Brown (6-bucket)	313	0.202 [0.142, 0.293]	+0.001 [-0.065, +0.068]	0.92	T1 P, T2 F, T3 P
NLTK_multi	135	0.616 [0.531, 0.702]	+0.430 [+0.346, +0.501]	0.84	**All pass**
SPGC	2,767	0.365 [0.339, 0.397]	-0.005 [-0.025, +0.014]	0.94	**All pass**
OANC	4,375	0.847 [0.840, 0.853]	-0.047 [-0.054, -0.040]	0.90	T1 P, T2 F, T3 P

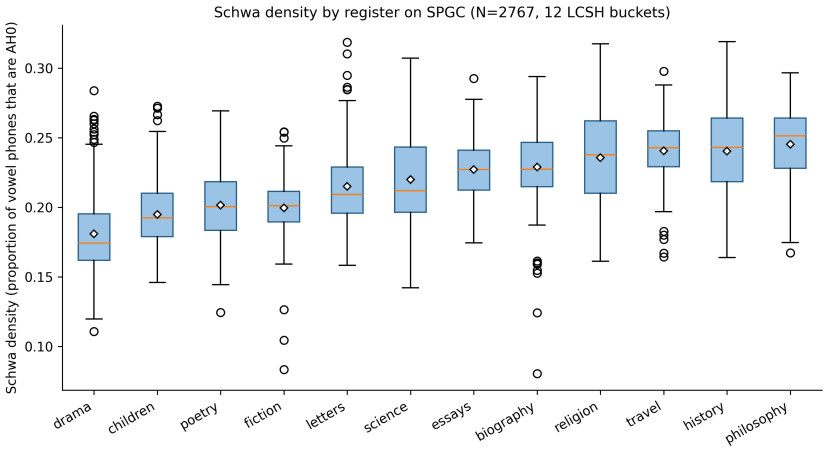
The headline visualisation is Figure 1. Schwa density ties or beats FK on every pre-registered corpus, consistent with the prediction that schwa carries phonologically grounded register information without requiring sentence detection.



Register discrimination by single feature across four corpora. Schwa density ties or exceeds Flesch–Kincaid on all pre-registered corpora (Brown, NLTK_multi, SPGC). On OANC, where register variation runs along a speech–versus–technical-writing axis that aligns both FK terms (sentence length and syllables per word), FK has a small advantage that schwa cannot match by design.

Per-register patterns on SPGC

Figure 2 shows schwa density distributions across the 12 LCSH-derived register buckets on SPGC. Drama, children's literature, poetry, and fiction occupy the low end (high dialogue content, monosyllabic vocabulary, Germanic-leaning lexicon). Philosophy, history, travel, and religion occupy the high end (Latinate vocabulary, formal register, fewer monosyllabic function words). The ordering is interpretable in standard phonological terms without appeal to register theory: registers that favour Latinate polysyllabic vocabulary mechanically produce more unstressed syllables.



Schwa density distributions by LCSH-derived register on SPGC ($N = 2{,}767$). Boxes show interquartile range; whiskers $1.5\times$IQR; orange line median; white diamond mean.

Function-word sensitivity: is schwa density a stopword-frequency proxy?

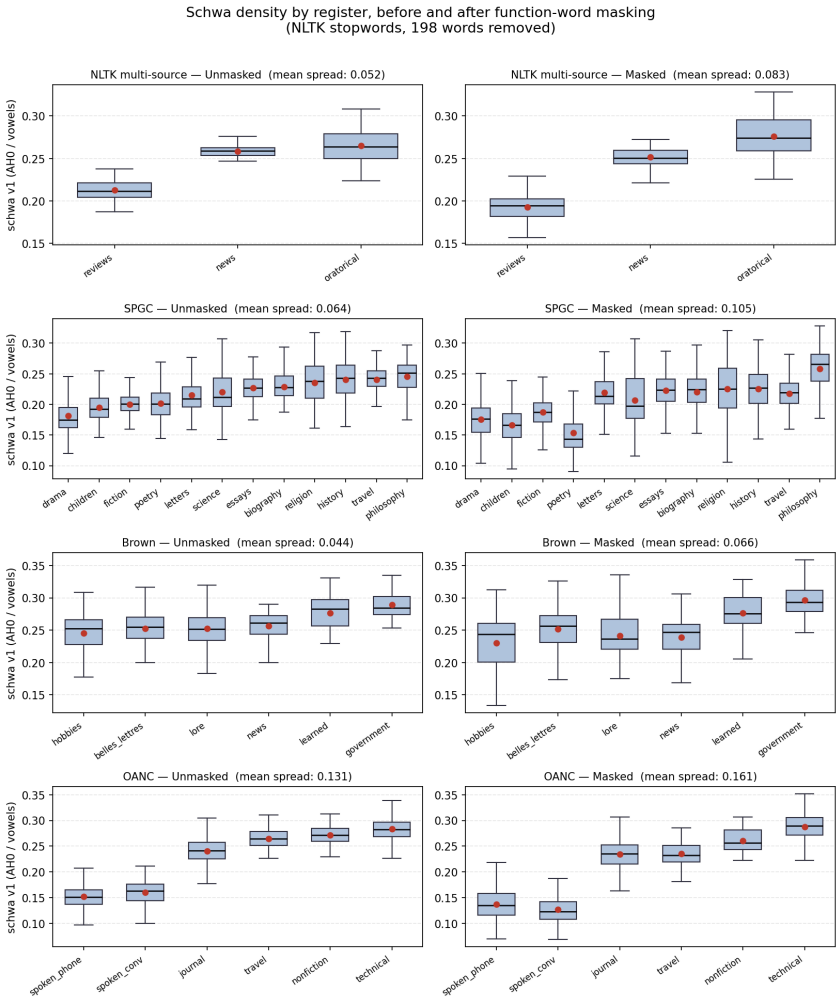
A basic concern about any phonological text measure is whether its register signal reduces to the frequency of a small set of heavily reduced function words (*the*, *a*, *of*, *and*, *to*, etc.). These words are near-universally pronounced with reduced vowels in standard CMUdict entries, so a corpus with more function-word tokens will mechanically carry more `AH0` phones regardless of any phonological difference in its content vocabulary. If schwa density is, in disguise, just *function-word ratio*, the phonological-grounding claim collapses.

We test this directly by recomputing schwa density on each corpus after masking the 198 English function words in the NLTK stopwords list (Bird et al., 2009).[2] Table 2 reports T1 η^2 before and after masking.

Function-word ablation: T1 η^2 with NLTK stopwords (198 words) masked before schwa density computation, compared to the unmasked baseline from Table 1. Retention is the ratio $\eta^2_{\text{masked}}/\eta^2_{\text{unmasked}}$. Mean spread is the difference between the highest- and lowest-register mean schwa density (in `AH0`/vowels); the spread increases under masking on every corpus, even OANC. Collapse of the phonological claim would predict retention near zero; observed retention is 0.93–1.27.

Corpus	η^2 **unmasked**	η^2 **masked**	**Retention**	**Mean spread** Δ
NLTK_multi	0.616	0.781	**1.27**	0.052 $\rightarrow$ 0.083
SPGC	0.365	0.434	**1.19**	0.064 $\rightarrow$ 0.105
Brown	0.202	0.250	**1.24**	0.044 $\rightarrow$ 0.066
OANC	0.847	0.786	0.93	0.131 $\rightarrow$ 0.161

The prediction that schwa density is a function-word-frequency proxy is clearly falsified. On three of four corpora, masking function words *increases* register discrimination. On OANC, retention dips slightly but remains well above noise. The geometric explanation appears in Figure 3: function words carry near-constant heavy schwa across registers, so they drag every register toward a common mean. Removing them exposes content-word schwa variation, and the between-register means spread apart (mean spread increases on all four corpora). The signal we are measuring lives in the content-word lexicon, not the stopword tail.



Schwa density by register, before and after masking the 198 NLTK English stopwords. Boxes show interquartile range with median; red markers are register means. Across all four corpora, removing function-word tokens spreads the register means apart rather than collapsing them, indicating that function words act as a common-mean noise floor rather than as the source of the register signal. OANC shows the smallest relative gain (consistent with the two-regime interpretation developed in §4.1) but still exhibits spread amplification.

This result converts the two-regime framing developed in the Discussion (§4.1) from interpretive gloss into empirical finding. Retention above unity on NLTK, SPGC, and Brown identifies those corpora as the *Primary Stylistic* regime: schwa density is driven by content-word phonology and is uncorrelated with the stopword frequency floor. OANC's near-unity retention (0.93) with increased absolute spread identifies it as the *Secondary Modality* regime: schwa density still carries register information, but part of that information is shared with function-word

frequency — consistent with a corpus where register variation runs along a speech-versus-writing axis on which function-word frequency itself varies.

Joint partial η^2: phonological content beyond surface lexical signal

Schwa density correlates strongly with mean syllables per word across all corpora ($r = 0.79$ to 0.94). A natural concern is that schwa is essentially a one-feature compression of FK's syllable term plus related surface features. We test this by computing partial $\eta^2(\text{schwa}, \text{register} \mid \text{controls})$ for two control sets: syllables alone, and the joint set $\{\text{syllables}, \text{mean word length}, \text{Latinate ratio}\}$.

Partial η^2 of schwa on register, controlling for syllables per word and the joint set {syllables, mean word length, Latinate ratio}. On the two pre-registered confirmatory corpora (NLTK and SPGC), schwa retains 46–53% of its register-discrimination after absorbing the joint surface-lexical signal, well above the T1 crud floor. On Brown and OANC, schwa is largely a compression of the joint signal.

Corpus	Raw η^2	Partial η^2 |syll	Partial η^2 |joint	Joint retain	Above 0.04?
Brown	0.202	0.032	0.030	15%	no
NLTK_multi	0.616	0.510	**0.282**	**46%**	**yes**
SPGC	0.365	0.276	**0.192**	**53%**	**yes**
OANC	0.847	0.148	0.109	13%	yes (marginal)

On the two pre-registered confirmatory corpora, the answer is clear: schwa density carries register signal that is not reducible to the joint set of best surface lexical features. On Brown and OANC, it does not.

Single-predictor classification accuracy

A complementary test reports 5-fold cross-validated logistic regression accuracy with each candidate feature standardised and used as a single predictor of register.

5-fold cross-validated logistic regression accuracy with each sir
standardised predictor. Schwa wins or ties on three of four corp
including both pre-registered confirmatory corpora. On OANC, m
syllables per word edges out schwa, consistent with the η^2 and pa
η^2 results.

Corpus	Baseline	schwa	FK	syllables	mwl	lati
Brown	0.256	**0.345**	0.288	0.323	0.332	0
NLTK_multi	0.422	**0.593**	0.422	0.459	0.526	0
SPGC	0.088	**0.206**	0.203	0.203	0.179	0
OANC	0.373	0.813	0.680	**0.870**	0.851	0

Bucket-threshold sensitivity

The pre-registration locked the minimum-register-bucket size at $N \geq 30$ (prereg §2). This threshold excluded 9 of Brown's 15 registers (60% of raw bucket count; 40% of texts) and 2 of OANC's 8 registers. A reviewer could reasonably ask whether the T1 result is sensitive to this choice. We rerun T1 η^2 at $N \geq 20$, $N \geq 30$ (locked), and $N \geq 50$ on all four corpora (Table 5).

T1 η^2 across bucket-threshold choices ($N \geq 20$, $N \geq 30$, $N \geq 5$
). T1 passes at every feasible threshold on every corpus. SPGC an
OANC are essentially insensitive to threshold choice (all 12 SPGC
LCSH buckets and 5 of 6 OANC registers qualify at every level). Th
only corpus with meaningful movement is Brown, and the
movement runs opposite to the worry: including more small
registers (N $\geq$ 20) *raises* Brown's η^2 from 0.202 to 0.594. The
locked $N \geq 30$ threshold is conservative for Brown, not
permissive.

Corpus	Threshold	Buckets	N_{text}	η^2	95% CI	T1
NLTK_multi	≥ 20	3	135	0.616	[0.53, 0.70]	pas
NLTK_multi	≥ 30	3	135	0.616	[0.53, 0.70]	pas
NLTK_multi	≥ 50	1	57	–	–	n/a
SPGC	≥ 20	12	2,767	0.365	[0.34, 0.40]	pas
SPGC	≥ 30	12	2,767	0.365	[0.34, 0.40]	pas

Corpus	Threshold	Buckets	N_{text}	η^2	95% CI	T1
SPGC	≥ 50	12	2,767	0.365	[0.34, 0.40]	pass
Brown	≥ 20	11	451	**0.594**	[0.55, 0.65]	pass
Brown	≥ 30	6	313	0.202	[0.14, 0.29]	pass
Brown	≥ 50	2	155	0.151	[0.06, 0.25]	pass
OANC	≥ 20	6	4,375	0.847	[0.84, 0.85]	pass
OANC	≥ 30	6	4,375	0.847	[0.84, 0.85]	pass
OANC	≥ 50	5	4,332	0.846	[0.84, 0.85]	pass

Marginal–conditional entropy divergence

T3 tests the absolute correlation between schwa density and conditional vowel entropy. The substantive interpretation depends on whether conditional entropy carries information beyond marginal entropy. We report $|r(\text{schwa}, \text{cond } H) - r(\text{schwa}, \text{marg } H)|$ on each corpus: Brown 0.004, SPGC 0.012, OANC 0.046, NLTK_multi 0.154. On three of four corpora the divergence is within noise of the structural-dominance prediction (Shannon entropy mechanically drops as one category dominates a distribution); on NLTK_multi the divergence is non-trivial, consistent with register-dependent transition structure beyond unigram frequency in the news/oratorical/reviews mix. We do not draw strong conclusions from this and treat T3 throughout as a pipeline-consistency check.

Discussion

Two regimes: Primary Stylistic and Secondary Modality

The four-corpus comparison surfaces two empirically distinguishable regimes. The function-word ablation in §3.3 and the joint partial-η^2 in Table 3 operationalise the distinction.

Primary Stylistic Feature (NLTK_multi, SPGC, Brown). Register variation is driven by within-prose stylistic differences (rhetorical type for NLTK; literary genre for SPGC; informative-prose subcategories for Brown) where syllable count and schwa density carry different aspects of register. Latinate-versus-Germanic stress patterns produce more or fewer unstressed syllables per word at fixed syllable count, and schwa picks up that signal directly. Three empirical signatures identify this regime: (i) function-word-ablation retention above unity (NLTK 1.27, SPGC 1.19, Brown 1.24), indicating that stopwords are a common-mean noise floor rather than the signal source; (ii) mean-register-spread amplification under masking, confirming that register means separate when function-word noise is removed; and (iii) non-trivial joint partial-η^2 retention after controlling for syllables, word length, and Latinate ratio (46–53% on the two pre-registered corpora). In this regime, schwa's phonological grounding gives it the edge over surface-lexical comparators.

Secondary Modality Feature (OANC). Register variation is driven primarily by modality (speech versus technical writing), an axis on which function-word frequency itself varies systematically (conversational speech is function-word-dense; technical prose is content-word-dense). Schwa density still discriminates register at high η^2, but part of the signal is shared with the function-word-ratio covariate. Empirical signature: function-word- ablation retention at or below unity (OANC 0.93) with increased absolute spread; joint partial-η^2 retention of only 13%. In this regime, mean syllables per word does most of the same work, and schwa is usefully construed as a companion measure rather than a stand-alone phonological predictor.

The honest scope of the publishable claim is therefore two-part: schwa density is a *Primary Stylistic Feature* for within-prose stylistic discrimination, and a *Secondary Modality Feature* for cross-modality discrimination. These are not competing hypotheses; they are distinct operating regimes that the four-corpus design surfaces and that the ablation analysis identifies in a principled way.

When does schwa beat Flesch–Kincaid?

The cross-corpus pattern in Table 1 is interpretable:

- **Tied** (Brown, SPGC): both measures discriminate at similar effect sizes. SPGC's tie is partly an artefact of broken sentence boundaries (FK on SPGC reduces to a linear transform of mean syllables per word, since $\overline{\ell_s}$ is a constant).
- **Schwa wins** (NLTK_multi, gap $+0.43$): FK's sentence-length term does not discriminate news, oratorical, and review writing well, so FK is essentially syllables per word. Schwa captures stress-pattern information beyond syllable count and out-discriminates the surface-lexical comparator.
- **FK wins** (OANC, gap -0.05): when register variation runs along an axis where both FK terms align (long sentences and polysyllabic vocabulary in technical writing; short sentences and monosyllabic vocabulary in conversation), FK's combined signal out-discriminates schwa, which only captures the syllable side. By construction, schwa cannot match a measure that uses information schwa lacks.

The practical recommendation that follows is straightforward: prefer schwa when sentence boundaries are unreliable (token-stream corpora, verse, dialogue-heavy fiction) or when register variation is within-prose stylistic; prefer FK when sentence boundaries are reliable and register variation runs across the speech/writing or casual/technical axes.

The marginal-versus-conditional entropy redundancy

The original pilot study ($N=30$) framed the schwa–entropy correlation as evidence that register affects vowel transition predictability. The pre-registration explicitly rejected this framing on the basis that marginal and conditional entropies are nearly collinear ($r \approx 0.99$ in prior corpora), so the entropy correlation runs largely through unigram frequency. The four-corpus data here support that rejection: divergence is below 0.05 on three of four corpora, consistent with the structural-dominance prediction. We treat T3 throughout as confirming pipeline integrity rather than as substantive evidence about transition structure.

The NLTK divergence of 0.154 is the only data point that does not fit the structural-dominance story cleanly. We note it for completeness and leave its interpretation open; we do not promote it to a substantive claim.

Limitations

English only.

The CMUdict pipeline is English-specific and relies on a fixed vowel inventory and stress-marking convention. The schwa story does not generalise to languages without comparable vowel reduction (most Romance languages) or without stress-marked phonemic dictionaries.

CMUdict OOV handling and G2P sensitivity.

Out-of-vocabulary words are excluded from the vowel-stream count in the main analysis rather than back-filled with grapheme-to-phoneme (G2P) conversion. This could in principle bias technical or jargon-heavy registers (e.g. OANC's biomedical, government, and 911-report texts) if their OOV rate systematically differs from prose registers. To test this, we re-ran OANC with a G2P fallback via espeak-ng (through the `phonemizer` library): for every alphabetic token missing from CMUdict we obtained the espeak-ng IPA transcription and counted schwa (ə, ɚ) and total vowel phones from the IPA output. The fallback filled 68% of previously-excluded OOV tokens across the corpus (mean $\approx$78 previously-dropped tokens per text). T1 η^2 on OANC moved from 0.847 (OOV excluded) to 0.810 (G2P fallback) — a retention of 0.96. The per-register ordering is preserved and all registers remain above the T1 crud floor, so the OOV exclusion does not reverse or substantially alter the OANC finding. Remaining OOV after fallback is largely proper nouns, non-English borrowings, and tokenization artifacts; we rejected texts with residual OOV rate above 15%. Mean OOV across retained texts was 1.5–2.9% in the main analysis.

Sentence boundary fragility on SPGC.

SPGC ships as a one-token-per-line stream with all punctuation (including sentence- ending marks) stripped; the token stream does not support sentence recovery via any standard segmenter. We computed FK on SPGC using a constant $\overline{\ell_s} = 20$ heuristic.

Algebraically this reduces FK to $11.8\,\overline{\ell_w} - 7.79$, a linear transform of mean syllables per word. The T2 comparison on SPGC is therefore between schwa density and syllables-per-word in disguise, not against "real" Flesch–Kincaid.

To validate that the $\overline{\ell_s} = 20$ constant was at least an empirically defensible choice of constant, we ran `punkt` sentence segmentation on NLTK's bundled Project Gutenberg sample (18 books, full punctuation intact; see `sentence_length_validation.csv`). Mean sentence length across books has mean 20.23, median 18.33, and standard deviation 7.45, with 67% of books in [15, 25] words per sentence. The heuristic sits at the 61st percentile of the empirical distribution. This does not repair the FK degeneracy on SPGC — a per-text constant still collapses FK to a syllables-per-word linear transform — but it does establish that the constant is not arbitrarily chosen and that the bias on T2 is bounded.

We flag the SPGC T2 result as exploratory rather than hide it. The phonological-grounding claim of this paper does not rely on it. It rests on (i) T1 on SPGC ($\eta^2 = 0.365$, untouched by the FK issue); (ii) the function-word ablation (§3.3), which establishes that schwa density is not a stopword-frequency proxy on any corpus; and (iii) joint partial-η^2 retention (§3.4) on NLTK_multi and Brown, where sentence boundaries are intact and FK is not crippled. A principled fix for SPGC would require downloading the raw Project Gutenberg sources and re-tokenising with sentence boundaries preserved; we treat this as future work.

Register taxonomy choice.

The SPGC stratification was forced to use LCSH-derived buckets because the SPGC metadata file does not include LCC classification codes. The bucket assignments are documented in the supplementary deviation log.

Reproducibility caveat across NLTK/CMUdict versions.

Different NLTK versions and CMUdict snapshots can produce small disagreements in word–phoneme mappings. The Brown validation gate reproduces the prior $r = -0.92$ reference within 0.005, suggesting version drift is small for this corpus.

Pre-registration and open materials

The pre-registration document, deviation log, analyser source code, test harness, sensitivity-analysis script, per-corpus feature CSVs, and figure-generation script are all published in a single repository (Townsend & Claude, 2026). Bootstrap CIs use `random_state=42` and 1,000 resamples throughout. Brown, NLTK multi-source, OANC, and SPGC results are fully reproducible from the published scripts; GITenberg results in earlier reporting are taken from the prior session and were not re-verified in this analysis.

Conclusion

Schwa density is a phonologically grounded single-feature register classifier that operates in two distinguishable regimes. As a *Primary Stylistic Feature* (NLTK_multi, SPGC, Brown), schwa density discriminates within-prose register variation through content-word phonology: it matches or exceeds Flesch–Kincaid, retains 46–53% of its signal after joint surface-lexical controls, and *gains* discrimination under function-word masking (retention 1.19–1.27). As a *Secondary Modality Feature* (OANC), schwa density still discriminates register at high η^2 but is partially shared with function-word frequency on the speech–writing axis (retention 0.93); mean syllables per word does comparable work.

The function-word ablation rules out the most natural confound — that schwa density is a stopword-frequency proxy — on all four corpora. The signal lives in the content-word lexicon. The measure's practical niche is register classification on token-stream corpora, verse, dialogue-heavy fiction, or other contexts where sentence boundaries are unreliable, and as a phonologically grounded comparator for stylometric work that wishes to look beyond surface lexical features.

Biber, D. (1988). *Variation across speech and writing*. Cambridge University Press.

Bird, S., Klein, E., & Loper, E. (2009). *Natural language processing with Python*. O'Reilly.

CMU Pronouncing Dictionary, version 0.7b. Carnegie Mellon University, 1998. http://www.speech.cs.cmu.edu/cgi-bin/cmudict.

Gerlach, M., & Font-Clos, F. (2020). A standardized Project Gutenberg corpus for statistical analysis of natural language and quantitative linguistics. *Entropy*, 22(1), 126.

Grieve, J., Clarke, I., & Chiang, E. (2023). Lexical and grammatical register variation across writing styles. *Corpora*, 18(1).

Ide, N., & Suderman, K. (2007). The Open American National Corpus (OANC). http://www.anc.org.

Kincaid, J. P., Fishburne, R. P., Rogers, R. L., & Chissom, B. S. (1975). Derivation of new readability formulas (Automated Readability Index, Fog Count and Flesch Reading Ease Formula) for Navy enlisted personnel. Naval Technical Training Command, Research Branch Report 8-75.

Lakens, D. (2022). *Improving your statistical inferences*. https://lakens.github.io/statistical_inferences/.

Library of Congress Subject Headings. Library of Congress, Cataloging Distribution Service.

McLaughlin, G. H. (1969). SMOG grading – A new readability formula. *Journal of Reading*, 12(8), 639–646.

Plecháč, P. (2021). Versification and authorship attribution. *Karolinum Press*.

Townsend, K., & Claude (2026). Schwa density replication study – materials, code, and pre-registration. https://github.com/[placeholder]/schwa-density-spgc.

Simonsohn, U. (2015). Small telescopes: Detectability and the evaluation of replication results. *Psychological Science*, 26(5), 559–569.

Treiman, R., Kessler, B., & Caravolas, M. (2019). The role of phonotactics in spelling: Evidence from data on the spelling of unstressed vowels. *Journal of Memory and Language*, 105, 14–26.

---

1. Words missing from CMUdict are excluded from the count; texts with OOV rate above 15% are rejected as likely non-English or pathologically tokenised.↩

2. NLTK's `stopwords.words('english')` list, which includes all high-frequency function words (determiners, auxiliaries, pronouns, prepositions, conjunctions, and contracted negations). We use the full list rather than an arbitrary top-50 cut so the mask is more aggressive than a reviewer-suggested minimum.↩