

StyleBot: the full collection

Author-balanced discipline benchmarks • 9 September 2026

The analysis now uses the usable main text across the supplied collection, supplemented by 44 acquired scholarly works. It replaces the 32-passage pilot with 146 works and approximately 1.15 million words. The calculated profiles below are provisional scholarly references. Their breadth is much stronger; precise national percentages and human-validated sentence accuracy have not been established.

What was analyzed

Collection	Works	Prepared main-text words
Supplied originals: 100 usable PDFs	102	721,582
External expansion	44	441,885
Total before sentence-level scoping	146	1,163,467

All 101 supplied PDFs are accounted for. One PDF contains three separately attributed contributions; one is excluded because the visible text is clipped. A Word file is an alternate format of an already counted work, not an additional essay. The final source-corrected sentence frame contains 1,145,017 word tokens in 40,824 candidate sentence units: 37,392 classified, 1,110 fragments and 2,322 uncertain. Prepared-word totals and final word tokens use different counting stages; quotation exclusions and source corrections also reduce the final frame.

Discipline	Original works	Added works	Total works	Author networks	Largest author weight
Literary criticism	39	10	49	25	4.0%
History	24	3	27	21	4.8%
Social research	22	11	33	25	4.0%
Legal scholarship	17	20	37	26	3.8%

The substantive result

Complex sentences account for about half of classified prose in these references. Literary and legal writing show higher estimated complex-sentence shares than history and social research. History and social research show higher simple-sentence shares. The overlapping ranges do not support fine rankings. “Complex” here is a grammatical category; it does not measure how sophisticated an idea is.

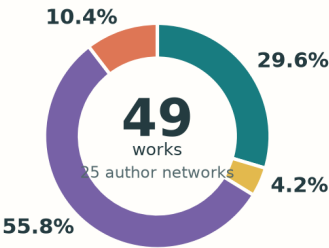
The largest composition change is legal writing. Its source-corrected complex-sentence estimate shifts from 61.7% in the supplied works to 56.5% after expansion, using the same author-network weighting. This illustrates why many sentences from a small number of writers are insufficient as a discipline benchmark.

Discipline charts: best current estimates

Author-balanced sentence distributions

Provisional reference • Classified units only • 146 scholarly works

Literary criticism



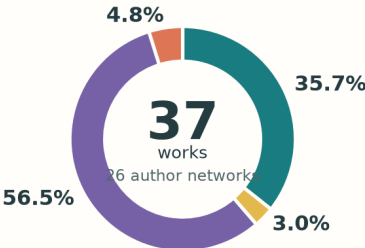
History



Social research



Legal scholarship



SimpleCompoundComplexCompound-Complex

Percentages are author-balanced averages among classified sentence units. These are descriptive references, not targets for students.

Discipline	Simple	Compound	Complex	Compound-complex
Literary criticism	29.6%	4.2%	55.8%	10.4%
History	39.3%	4.7%	49.7%	6.3%
Social research	40.8%	4.5%	48.5%	6.2%
Legal scholarship	35.7%	3.0%	56.5%	4.8%

The charts use the source-corrected primary parser estimates. An essay first contributes its own distribution; a longer essay does not receive extra weight. Displaying a decimal reports the calculation, not accuracy to a tenth of a percentage point. Rounding can cause a displayed row to differ slightly from 100%.

How the weighting prevents domination

1. Calculate each essay’s four proportions using only its classified sentence units. Keep fragments and uncertain units visible in separate coverage counts.
2. Join works connected by a shared author, including coauthors and confirmed name variants. This creates an author network. It is a dependence-control device, not a claim that all coauthors write alike.
3. Give each network equal weight within its discipline. Divide that network’s weight equally among its essays. With K networks and D essays in a network, each essay receives 1/(K × D). Every usable essay remains included.

For example, a literary author network with seven essays shares one 4% contribution. Its individual essays each receive about 0.57%, whether an essay has 100 classified sentences or 1,000. The legal network containing ten Laurence Tribe works receives 3.85% of the expanded legal reference; its ten works do not become ten independent author votes. Full exposure includes every coauthored work, without dividing away a coauthor’s exposure.

Observed-author resampling ranges

Discipline	Simple	Compound	Complex	Compound-complex
Literary criticism	26.2–33.1%	3.5–4.8%	52.8–58.8%	8.5–12.5%
History	35.8–42.8%	3.7–5.8%	46.4–53.2%	4.7–7.9%
Social research	37.0–44.3%	3.5–5.6%	45.4–51.8%	4.9–7.8%
Legal scholarship	32.6–38.9%	2.3–3.7%	53.4–59.5%	3.9–5.8%

These are marginal central 95% ranges from 5,000 resamples of whole observed author networks. They describe variation under that resampling model. They do not estimate national sampling error, source-selection bias or classifier error, and the four endpoints are not one simultaneous composition.

Discipline	Simultaneous stability radius	Classified coverage, weighted
Literary criticism	±3.67 points	89.5%
History	±3.80 points	94.6%
Social research	±3.79 points	93.7%
Legal scholarship	±3.34 points	93.3%

The simultaneous radius is the 95th percentile of the largest absolute change among the four shares. All fields exceed the prespecified 3-point stability tolerance. Coverage remains below the 95% weighted threshold, with literary writing the weakest. These are diagnostic engineering tolerances, not proof of educational effectiveness or nationally correct ratios.

Validation: what the checks actually establish

The grammar rules and original prepared input frame were frozen before drawing 400 probability-core units, 100 per discipline, spanning all 146 works and positions across their main text. Another 100 difficult units were kept as a separate challenge set. Forty randomly chosen core units received an independent second review, with 49 additional targeted reviews of uncertain cases: 38 core units and 11 challenge units. The 89 independently rechecked units remain distinct from the 500 initial labels. Initial reviewers were blind to parser predictions. Reviewers were AI agents, not human annotators; source images were checked for targeted issues, not independently for every sampled unit.

Discipline	spaCy exact label agreement	Stanza exact label agreement	Audit distribution discrepancy
Literary criticism	85.0%	79.0%	3.74 points
History	90.5%	91.9%	2.09 points
Social research	87.1%	86.9%	2.77 points
Legal scholarship	89.5%	88.3%	2.04 points

These weighted figures use the 400-unit probability core and adjudicated reference labels against the original frozen text. “Exact label” includes both the four types and fragment/uncertain status, so abstentions do not disappear. The discrepancy compares conditional four-type distributions within the audit, with each rater’s classified denominator; it is a diagnostic point estimate, not a correction applied to the whole corpus.

Initial labels and disagreements remain preserved. One targeted literary adjudication changed an uncertain label to compound-complex after checking the PDF; the initial literary exact agreement was 83.7%, before adjudication. The source-corrected full corpus has not had a new untouched validation sample. A small change in its aggregate percentages does not erase the source errors found in the original audit.

Sensitivity to analytical choices

Discipline	Second parser	27 source corrections	Quoted units retained
Literary criticism	2.09 points	0.04 points	0.83 points
History	1.34 points	0.01 points	0.25 points
Social research	0.25 points	0.01 points	0.12 points
Legal scholarship	1.06 points	0.04 points	0.24 points

Each cell is the largest absolute shift among four shares from the same frozen baseline with essay weights held fixed. The quotation scenario restores 1,029 standalone quoted sentence units from otherwise eligible paragraphs; it does not restore all block quotations removed earlier. Small net differences can conceal individual errors or shared parser mistakes. The second parser is a sensitivity check, not an accuracy certificate.

Corpus classification used local spaCy 3.8.16 / en_core_web_trf 3.8.0 and Stanza 1.14.0 English EWT, separately from production StyleBot’s Astra integration. The rules distinguish finite clauses from phrases; shared-subject predicates remain one clause.

Exclusions and source corrections

Scope	Treatment and reason
Entire source: S058	Rural Puerto Rico in the Early Twentieth Century Reconsidered: Land and Society, 1899–1915. The right edge is visibly clipped; the publisher replacement has the same defect. Excluded rather than reconstructing missing words.
S047: Debating Patriarchy	Three attributable contributions retained as separate works: Julia Adams, Benita Roth and Pavla Miller. Unattributed introductory material excluded.
Supplementary Word document	Alternate format of the Color Money source (S070), with 719 extractable words and three embedded images; not counted as another independent work. Full equivalence of image content was not established.
External candidate L018	Complete text was not obtained from the official repository. Not counted. L025 is a different accessible work by the same author, explicitly identified as such.
Apparatus and reproduced material	Titles/bylines, abstracts, navigation, running headers/footers, notes, references, tables, captions and displayed source quotations excluded. Inline quotations embedded in an author’s own sentence generally retained; transcript speech and reproduced verse excluded.
Source defects found after freezing	27 source-confirmed corrections across 22 works: footnote markers, heading/caption contamination, split/merged boundaries, damaged extraction and a reproduced ballad block. Separate scenario replaces 90 units with 39; original audit remains intact.
Incomplete or ambiguous units	Source lead-ins and unresolved grammar remain uncertain or fragment. They are reported outside the four-type denominator, not silently forced into a category.

The separate source ledger accounts for each of the 101 original PDFs and each of the 44 external works, including attribution, discipline, prepared words, page ranges or source locations, hashes and exclusion reasons. Line-level exclusions and later paragraph exclusions are kept separate; they cannot be added as though they were disjoint word counts. Initial preparation and post-audit corrections are also separated. Documented repairs do not mean every extracted word was manually validated.

External research followed the original collection

The 44 additions comprise 10 literary works, 3 historical works, 11 social-research works and 20 legal works. Selection sought new author networks and complete scholarly argument from publisher or university sources, without consulting sentence-type shares. Additions include Victorian and Romantic criticism, historical case studies, sociological research, sociolegal work and American legal argument. Genre and publication differences remain visible; four empirical sociolegal works are classified by their research task as social research.

The sources are purposively selected and not a random national sample. Some additions share journals or themed volumes. Available publication/topic deletion checks are useful but do not cover publication metadata across the entire original collection. Published scholarly writing is the defensible label; faculty rank has not been independently verified for every author.

What this means for StyleBot

Use these estimates as broad discipline references with transparent coverage and uncertainty. The expansion addresses author concentration. It does not yet establish the fine percentage accuracy needed for rigid scoring. All four fields pass the 20-network breadth floor and 5% author-exposure cap; all still fail one or more coverage, stability or measurement checks.

Priority	Next improvement
1 • Repair measurement	Use the documented extraction and boundary failures to improve processing, especially literary quotation, titles, parentheticals and source damage. Freeze a new version and evaluate it on a new untouched sample. Preserve the current audit as the baseline.
2 • Calibrate with educators	Have grammar-trained reviewers independently label a stratified sample, resolve disagreements with a published clause convention, and report human agreement separately. This is a validation study, not a new approval gate for the educator.
3 • Expand strategically	After measurement repairs, add new author networks from underrepresented journals, genres, periods and institutions, following the collection’s scholarly purpose. Recompute after each batch; do not choose texts because they move percentages toward a preferred answer.
4 • Teach choices, not quotas	Let a student compare their own writing, inspect sentence functions, revise one passage, and explain its intended effect to a PLF. Track the student’s ability to transfer that choice to a new independent writing task.

For planning only, scaling the observed network variation suggests approximately 13 more literary, 13 historical, 15 social-research and 7 legal networks to approach a 3-point simultaneous resampling radius: 48 additional networks in total. This assumes comparable future variation and does not repair measurement error or guarantee reliability. It is a projection, not a claim that those additional works have been acquired.

A student should be able to ask: Where would a short independent sentence sharpen my claim? Where does a dependent clause make a qualification necessary? What changes if I connect two claims rather than separate them? A PLF can review the student’s reasoning and revision. Matching a benchmark pie chart should not determine a grade or count as demonstrated growth.

Deliverables and traceability

The accompanying files are: StyleBot-full-corpus-benchmark-review.md (this report), StyleBot-source-exclusions.md (source-by-source ledger), StyleBot-corpus-expansion-review.md (selection details and linked bibliography), StyleBot-weighted-benchmark-data.json (derived discipline and essay counts, weights and hashes), and StyleBot-discipline-benchmarks.png (standalone charts). Source texts, token ledgers and blind annotations remain in the private research workspace. No production app or deployment was changed by this research pass.

Primary methodological references

[spaCy English model documentation](#) and [Stanza dependency parsing](#) describe the parser families. [Universal Dependencies: clausal complements](#) documents the dependency distinction used by the mapping. [Field and Welsh, Bootstrapping clustered data](#) motivates cluster-aware resampling. [AAPOR Transparency Initiative](#) informs the disclosure of nonprobability-sample limitations. The exact external articles are linked in the expansion bibliography.