

FOR LEGAL / INVESTIGATIVE USE

FORENSIC ANALYSIS REPORT

MCRO | Children Tracking Groups Forensics — A1–F1 Part 2: Language & Terminology Divergence

Minnesota Court Records Online (MCRO) — Fourth Judicial District, Hennepin County

Report Subject	Language & Terminology Divergence Across 13 Tracking Font PDF Groups
PDF Groups Analyzed	A1, A2, A3, A4, A5, A6, B1, C1, D1, D2, D3, E1, F1 (13 groups)
Font Families	6 letter-families: A (TTF-A), B (TTF-B/CFF), C (TTF-C), D (TTF-D), E (TTF-E), F (TTF-F)
Total Documents in Scope	1,942 PDFs across 245+ cases
Documents with Term Strings	1,810 of 1,942 (93.2% coverage)
Total Term Occurrence Rows	100,555 across 319 distinct terms
Date Range of Filings	January 2017 – August 2025
Data Source	PostgreSQL forensic database (Supabase) — READ-ONLY analysis
Database Tables Used	children_tracking_groups, children_doc_strings, hashpack_rows, hashpack_rows_tags, children
Series	Part 2 of Font Tracking A1–F1 series (Part 1: Structural/Metadata/Signature Analysis)
Report Date	March 2026
Reproducibility	All findings are query-reproducible. SQL in Appendix.

IMPORTANT DISCLAIMER

This report presents objective forensic data findings reproducible from the live source database. All statistics were freshly computed against the live PostgreSQL/Supabase database (project `ibfmjtwahkwqzcmeyqii`) on the date of this report. No pre-computed values from prior reports or uploaded reference files have been used as factual inputs.

The analyses contained in this report do not assert conclusions about fraud, fabrication, criminal intent, or motive. Each finding documents a measurable pattern in the data. Causal explanations are not provided in Findings sections unless directly and unambiguously supported by the database evidence examined in this report.

This is Part 2 of the Font Tracking A1–F1 series. Part 1 established structural, metadata, and signature boundaries between the 13 tracking font groups. This Part 2 tests whether those structural boundaries are reinforced or contradicted by language and terminology patterns in the documents' extracted text.

Independent verification recommended. All SQL queries are provided in full in the Appendix. A credentialed digital forensics examiner (CFCE, EnCE, or equivalent) should independently verify these findings before use in formal proceedings.

EXECUTIVE SUMMARY

The 13 tracking font groups that Part 1 showed are structurally distinct document production pipelines also produce linguistically distinct documents. Language partitioning strongly reinforces structural partitioning, with the sharpest terminological boundaries aligning precisely with Part 1's structural "brick walls."

1. 319 distinct terms (string_norm) appear across the 13-group cohort's 100,555 occurrence rows in 1,810 documents with term strings (93.2% coverage of the 1,942-document corpus).
2. 101 terms (31.7% of all 319 terms) are exclusive to a single pdf_group — appearing in one group and zero others. B1 leads with 45 exclusive terms; C1 has 18; F1 has 16. Five groups (A1, A5, A6, D2, E1) have zero exclusive terms.
3. At the letter-family level, Family B holds 45 family-exclusive terms, Family A holds 19, Family C holds 18, Family F holds 16, Family D holds 5, and Family E holds zero — every Family E term also appears in other families.
4. The A6 → B1 boundary is the sharpest language discontinuity in the corpus: Jaccard term similarity = 0.005 (effectively zero overlap). A6 has only 1 term ("detention"); B1 has 219 — a vocabulary explosion of 21,800% at the same structural boundary Part 1 identified as the sharpest metadata/signature wall.
5. C1 → D1 is the highest-similarity boundary (Jaccard = 0.577), confirming shared Rule 20 competency pipeline vocabulary. B1 → C1 is second (Jaccard = 0.544). Both pairs share deep "experts," "protect," and "indefinite" terminology — the language of competency evaluation proceedings.
6. D3 is a linguistic outlier within the D-family: its top 5 terms are judicial officer names (Julia Dayton Klein, Danielle Mercurio, George Borer, Michael Browne) and "Google" — a vocabulary fingerprint entirely unlike the Rule 20 evaluation language dominating D1 and D2.
7. D2 has the lowest row-level text uniqueness in the corpus: only 753 distinct row hashes across 10,784 total rows (7.0% uniqueness), indicating extreme template reuse. D1 is second-lowest at 12.7%.
8. F1 (4 documents) has 16 exclusive terms including DSM-5 diagnostic language ("delusional disorder," "hallucination," "untreated mental illness") and behavioral descriptors ("pressured speech," "tangential," "required redirection") — a vocabulary density of 4.0 exclusive terms per document, the highest of any group.
9. Cross-group row hash sharing is concentrated on the B1–C1–D1 axis: C1 ↔ D1 share 1,681 identical text lines, B1 ↔ C1 share 1,019, and B1 ↔ D1 share 646. The A-family internal linkage (A2 ↔ A4: 1,288 shared lines) exceeds most cross-family sharing, confirming intra-family template reuse.
10. The language data produces a convergent signal: groups that Part 1 showed are structurally isolated (different fonts, different metadata pipelines, different signature architectures) are also terminologically isolated. The structural walls are also vocabulary walls.

Corpus Overview — Term Coverage by PDF Group

Group	Docs	With	Without	Total Rows	Distinct	Vocab	Excl	Excl %
-------	------	------	---------	------------	----------	-------	------	--------

	Strings		Strings		Terms	Richness	Terms	
A1	7	7	0	36	9	0.2500	0	0.0%
A2	65	65	0	775	83	0.1071	6	7.2%
A3	51	51	0	310	29	0.0935	1	3.4%
A4	9	9	0	512	104	0.2031	10	9.6%
A5	6	6	0	39	13	0.3333	0	0.0%
A6	1	1	0	2	1	0.5000	0	0.0%
B1	235	233	2	16,592	219	0.0132	45	20.5%
C1	1,089	991	98	45,227	204	0.0045	18	8.8%
D1	104	104	0	29,100	143	0.0049	4	2.8%
D2	131	131	0	6,319	31	0.0049	0	0.0%
D3	238	206	32	369	46	0.1247	1	2.2%
E1	2	2	0	170	38	0.2235	0	0.0%
F1	4	4	0	1,104	110	0.0996	16	14.5%

KEY METRICS

- 101 of 319 terms (31.7%) are exclusive to a single group — nearly one-third of the entire vocabulary is group-locked
- A6 → B1 Jaccard = 0.005 — the sharpest language boundary in the corpus, matching Part 1's sharpest structural wall
- D2 text uniqueness = 7.0% — 93% of D2's 10,784 text rows are duplicated (extreme template reuse)
- F1 exclusive density = 4.0 exclusive terms per document — highest in corpus, from only 4 documents
- B1 exclusive count = 45 terms — the largest exclusive vocabulary of any single group, including named examiners (Sonia Reardon: 90 occurrences across 35 docs)
- C1 ↔ D1 shared text lines = 1,681 — the strongest cross-group row-level linkage, confirming shared Rule 20 pipeline
- D3 top 5 terms are judicial officer names — the only group where named persons dominate the vocabulary, consistent with hearing-notice document function
- Zero exclusive terms in E1 — all 38 terms in this 2-document group also appear in other groups

METHODS

Cohort Definition

The analysis cohort is identical to Part 1: all 1,942 rows in `children_tracking_groups` where `pdf_group` is one of A1, A2, A3, A4, A5, A6, B1, C1, D1, D2, D3, E1, or F1. Document counts and group assignments were verified against the live database and match Part 1 exactly.

Tables and Join Keys

Primary analytical tables: `children_tracking_groups` (`pdf_group` assignment, join key: `evidence_uid`), `children_doc_strings` (term-level extracted text, join key: `evidence_uid`), `hashpack_rows` (row-level text with SHA256 hashes, join key: `evidence_uid`), `hashpack_rows_tags` (row classification tags), and `children` (document metadata context).

Term Exclusivity Definition

A term (`string_norm`) is “exclusive to `pdf_group` X” if it appears in at least one document in group X (via `children_doc_strings` joined to `children_tracking_groups` on `evidence_uid`) and appears in zero documents in any other `pdf_group` within the 13-group cohort. Family-level exclusivity applies the same logic at the letter-family aggregation (A*, B*, C*, D*, E*, F*).

Jaccard Similarity

For each pair of groups (or families), the Jaccard index is computed as $|\text{intersection}| / |\text{union}|$ of their distinct `string_norm` vocabularies. Values range from 0.0 (no shared terms) to 1.0 (identical vocabularies).

Language Shift Score

For each adjacent boundary pair (e.g., A6 → B1), the language shift score is defined as (terms exclusive to left group + terms exclusive to right group) / total distinct terms in the union of the pair. Higher scores indicate sharper vocabulary discontinuity.

Row-Level Text Overlap

Cross-group row hash sharing is measured by counting distinct `row_sha256` values from `hashpack_rows` that appear in documents belonging to two or more different `pdf_groups`. This measures literal text-line identity between groups, independent of term-level vocabulary analysis.

FINDINGS

Finding 1: Corpus Overview — doc_strings Coverage by PDF Group

Background

Before analyzing term distributions across groups, coverage must be established: how many documents in each group have extracted term data, and what is the volume and diversity of that data?

Findings

Of the 1,942 documents in the 13-group cohort, 1,810 (93.2%) have at least one `children_doc_strings` row. The 132 documents with zero term strings are concentrated in two groups: C1 (98 documents, 9.0% of C1) and D3 (32 documents, 13.4% of D3). All other groups have 100% or near-100% coverage.

The total cohort contains 100,555 term occurrence rows across 319 distinct terms (`string_norm` values) and 20 distinct `string_group_type` categories. C1 dominates with 45,227 rows (45.0% of all cohort rows), followed by D1 with 29,100 rows (28.9%).

Vocabulary richness ratio (distinct terms / total rows) varies dramatically: A6 has 1 term in 2 rows (0.500); D2 has 31 terms in 6,319 rows (0.005). This 100-fold range indicates fundamentally different document types — from brief boilerplate documents to dense forensic evaluation reports.

Significance

The 93.2% coverage rate confirms that term-level analysis captures the vast majority of the cohort. The two coverage gaps (C1 and D3) are not large enough to invalidate group-level statistics, but they are noted for completeness.

Connecting Context

Part 1 identified C1 as 56% of the cohort and predominantly Rule 20 competency proceedings. The term data confirms this: C1's 45,227 rows are dominated by "experts" (30.3%) and "protect" (21.4%) categories. D3's 13.4% coverage gap aligns with Part 1's finding that D3 consists predominantly of hearing notices — brief procedural documents with minimal extractable term content. See Appendix A1–A2.

Finding 2: Term Vocabulary Inventory — Size and Composition per Group

Background

Vocabulary size and composition reveal the functional purpose of each group's documents. A group producing detailed forensic psychiatric evaluations will have a larger, more specialized vocabulary than a group producing procedural hearing notices.

Findings

The vocabulary size table ranked by distinct term count:

Group	Distinct Terms	Total Rows	Vocab Richness (terms/rows)	Term Categories	Rank
B1	219	16,592	0.0132	19	1
C1	204	45,227	0.0045	18	2
D1	143	29,100	0.0049	17	3
F1	110	1,104	0.0996	15	4
A4	104	512	0.2031	15	5
A2	83	775	0.1071	15	6
D3	46	369	0.1247	13	7
E1	38	170	0.2235	14	8
D2	31	6,319	0.0049	12	9
A3	29	310	0.0935	10	10
A5	13	39	0.3333	8	11
A1	9	36	0.2500	6	12
A6	1	2	0.5000	1	13

B1 has the largest vocabulary (219 terms) despite being only 12.1% of cohort documents. C1, the largest group by document count (56.1%), has the second-largest vocabulary (204 terms). At the other extreme, A6 has a single term ("detention") in its lone document.

Significance

The vocabulary sizes cluster into four tiers: large (B1: 219, C1: 204, D1: 143, F1: 110, A4: 104), medium (A2: 83, D3: 46, E1: 38, D2: 31, A3: 29), small (A5: 13, A1: 9), and minimal (A6: 1). These tiers align with Part 1's document-function classifications.

Connecting Context

Part 1 showed B1 spans 27 filing types (the broadest range) and C1 concentrates in Rule 20 competency proceedings. Their vocabulary sizes reflect this: B1's breadth produces vocabulary breadth; C1's depth in one domain produces vocabulary depth. D2's extremely low richness ratio (0.005) matches Part 1's finding that 98% of D2 documents are Rule 20.01 evaluation orders — highly templated documents. See Appendix A3.

Finding 3: Group-Exclusive Terms — The Centerpiece Finding

Background

A term exclusive to a single pdf_group provides the strongest possible evidence of vocabulary isolation: the term exists within that group's documents and nowhere else in the 13-group cohort. The count and nature of exclusive terms reveal whether each group's vocabulary is a subset of a shared lexicon or contains genuinely unique language.

Findings

101 of 319 terms (31.7%) are exclusive to a single group. The remaining 218 terms (68.3%) appear in two or more groups. Five groups (A1, A5, A6, D2, E1) contribute zero exclusive terms. The distribution:

Group	Exclusive Terms	Occurrence Rows	Docs Affected	Group Vocab Size	Excl % of Vocab
B1	45	220	106	219	20.5%
C1	18	237	100	204	8.8%
F1	16	46	28	110	14.5%
A4	10	20	11	104	9.6%
A2	6	29	11	83	7.2%
D1	4	18	8	143	2.8%
A3	1	1	1	29	3.4%
D3	1	2	1	46	2.2%

Selected exclusive terms by group (representative examples from each group with exclusive terms):

Group	Term (string_norm)	Category	Occ Rows	Docs	Significance
B1	Sonia Reardon	examiner	90	35	Named examiner
B1	Kimberly Turner	examiner	16	2	—
B1	Amanda Jung	liason	10	5	—
B1	state-operated treatment	itinerary	8	4	—
B1	misconduct	fraud	7	4	—
B1	grandiose	dsm5	4	1	—
C1	Psy.D., LP, ABPP	experts	88	27	Credential string
C1	Amanda Powers	examiner	66	32	—
C1	Jennifer Flowers	examiner	20	7	—
C1	Gregory Hanson	examiner	18	7	—
C1	behavioral notes	control	4	4	—
F1	delusional disorder	dsm5	4	2	DSM-5 diagnosis
F1	hallucination	dsm5	4	1	Clinical symptom
F1	pressured speech	mirror	2	2	Behavioral obs.
F1	refusal to participate	mirror	4	2	—

A4	forensic navigator	monitoring	7	1	Statutory role
A4	controlling	control	3	1	—
D1	metadata	fraud	10	2	Pro se filing term
D1	informed consent	fraud	4	2	—
D1	Brady	fraud	2	2	Legal doctrine

Significance

The exclusive term inventory reveals three distinct vocabulary categories: (1) Named persons — examiners, attorneys, liaisons specific to certain pipeline segments (Sonia Reardon in B1; Amanda Powers in C1); (2) Clinical/diagnostic language — DSM-5 terms and behavioral descriptors concentrated in F1; (3) Legal doctrine terms — pro se filing language (“metadata,” “Brady,” “informed consent”) exclusive to D1. The vocabulary isolation is not random — it reflects distinct document authorship streams.

Connecting Context

Part 1 showed B1 spans the broadest filing type range (27 types) and the longest active period (8.5 years). Its 45 exclusive terms include named examiners and unique procedural terms consistent with a mature, diverse document pipeline. C1’s exclusive terms are dominated by specific forensic psychologist names and credential strings, consistent with Part 1’s finding that C1 concentrates in Rule 20 competency evaluation orders. F1’s 16 exclusive terms are entirely clinical/behavioral — consistent with detailed forensic psychiatric evaluation content in just 4 documents. See Appendix A4.

Finding 4: Cross-Group Term Sharing — Jaccard Similarity Matrix

Background

Jaccard similarity measures vocabulary overlap between groups. Values near 1.0 indicate nearly identical vocabularies; values near 0.0 indicate almost no shared terms. The 13×13 matrix reveals which groups share a common vocabulary and which are linguistically isolated.

Findings

Top 10 most-similar group pairs by Jaccard term similarity:

Rank	Group 1	Group 2	Shared Terms	Union Terms	Jaccard
1	C1	D1	127	220	0.577
2	B1	C1	149	274	0.544
3	B1	D1	110	252	0.437
4	A4	D1	72	175	0.411
5	C1	F1	86	228	0.377
6	D1	F1	68	185	0.368
7	A2	A4	50	137	0.365
8	B1	F1	84	245	0.343
9	A4	C1	76	232	0.328
10	A4	B1	78	245	0.318

Bottom 5 most-dissimilar group pairs: A6 ↔ B1 (0.005), A6 ↔ C1 (0.005), A6 ↔ D1 (0.007), A4 ↔ A6 (0.010), A2 ↔ A6 (0.012). All involve A6, which has only 1 term. Excluding A6, the most dissimilar pairs are A1 ↔ E1 (0.022) and A1 ↔ B1 (0.041).

Significance

The B1–C1–D1 triangle forms a linguistic cluster with Jaccard values between 0.44 and 0.58. These three groups share the deep vocabulary of Rule 20 competency proceedings: evaluation orders, expert reports, commitment language. F1 joins this cluster at a lower but still meaningful level (0.34–0.38). The A-family subgroups are linguistically isolated from most other families, with intra-family similarity (A2 ↔ A4: 0.365) exceeding most cross-family comparisons.

Connecting Context

Part 1 showed B1, C1, and D1 all carry judicial officer signatures and produce Rule 20-related filing types. The Jaccard similarity confirms they share not only structural features but also terminological content. The D3 group, which Part 1 identified as anomalous (zero judicial signatures, 88% hearing notices), shows low Jaccard values with its own D-family siblings D1 (0.235) and D2 (0.222) — reinforcing its outlier status. See Appendix A5.

Finding 5: “Brick Wall” Language Discontinuity Testing

Background

Part 1 identified sharp structural boundaries at each adjacent group transition. This finding tests whether those same boundaries correspond to vocabulary discontinuities by computing language shift scores and Jaccard similarity at each boundary.

Findings

Boundary	Left Terms	Right Terms	Excl Left	Excl Right	Union	Shift Score	Jaccard
A6 → B1	1	219	0	218	219	0.995	0.0046
B1 → C1	219	204	70	55	274	0.456	0.5438
C1 → D1	204	143	77	16	220	0.423	0.5773
D1 → D2	143	31	112	0	143	0.783	0.2168
D2 → D3	31	46	17	32	63	0.778	0.2222
D3 → E1	46	38	29	21	67	0.746	0.2537
E1 → F1	38	110	5	77	115	0.713	0.2870

The A6 → B1 boundary has a language shift score of 0.995 and Jaccard of 0.005 — effectively total vocabulary replacement. A6 contributes only 1 term (“detention”); B1 introduces 218 new terms. This is the sharpest language discontinuity in the corpus, precisely matching Part 1’s identification of A → B as the sharpest structural wall.

The D1 → D2 boundary has the second-highest shift score (0.783): D2 introduces zero new terms (all 31 D2 terms are already in D1), but 112 of D1’s 143 terms vanish. D2 is a strict vocabulary subset of D1 — consistent with D2 being a more templated version of the same document type.

The B1 → C1 and C1 → D1 boundaries have the lowest shift scores (0.456 and 0.423 respectively) and the highest Jaccard values — confirming these are the smoothest vocabulary transitions, consistent with shared Rule 20 pipeline vocabulary across these groups.

Significance

Language shift scores rank the boundaries in the same order as Part 1’s structural discontinuities: A → B is the sharpest wall in both analyses, and B → C → D1 is the smoothest transition in both. The convergence between structural and linguistic analysis is a strong signal: these boundaries are not artifacts of a single measurement dimension.

Connecting Context

Part 1’s boundary analysis showed: A → B: font format change (TTF → CFF), producer pipeline shift (Winnovative → Word), 0% → 99% judicial signatures. Part 2 now confirms: vocabulary replacement rate of 99.5% at the same boundary. The structural wall and the language wall are co-located. See Appendix A6.

Finding 6: string_group_type Taxonomy Distribution Across Groups

Background

The string_group_type column classifies each term occurrence into a semantic category (e.g., “experts,” “protect,” “dsm5,” “officer”). The distribution of these categories across groups reveals whether different groups produce different types of legal/clinical content.

Findings

The top 3 string_group_type categories by group (percentage of group’s total rows):

Group	#1 Category	#2 Category	#3 Category
A1	itinerary (72.2%)	dsm5 (13.9%)	protect (5.6%)
A2	itinerary (31.1%)	dsm5 (19.0%)	control (9.7%)
A3	itinerary (62.6%)	protect (9.7%)	dsm5 (7.7%)
A4	itinerary (23.8%)	fraud (12.9%)	protect (12.7%)
A5	itinerary (61.5%)	fraud (12.8%)	dsm5 (10.3%)
A6	itinerary (100%)	—	—
B1	experts (25.4%)	protect (21.0%)	indefinite (16.7%)
C1	experts (30.3%)	protect (21.4%)	indefinite (13.2%)
D1	experts (47.1%)	protect (16.3%)	itinerary (7.8%)
D2	experts (47.7%)	protect (16.5%)	itinerary (7.9%)
D3	officer (41.2%)	truth (14.6%)	protect (13.3%)
E1	protect (27.6%)	experts (23.5%)	indefinite (14.1%)
F1	dsm5 (21.8%)	experts (19.6%)	protect (13.1%)

Three distinct taxonomy profiles emerge: (1) The A-family is dominated by “itinerary” (detention/transport language), consistent with charging instruments. (2) B1, C1, D1, and D2 share an “experts + protect + indefinite” profile — the language of forensic psychiatric evaluation and commitment proceedings. (3) D3 is uniquely dominated by “officer” (41.2%) — judicial officer names appearing in hearing notices.

F1 is the only group where “dsm5” is the #1 category (21.8%), reflecting its concentration of clinical diagnostic language. No other group has “dsm5” in the top position.

Significance

The taxonomy distribution independently confirms the functional classification of groups established by Part 1’s filing type analysis: A-family = charging instruments (itinerary-heavy), B1/C1/D1/D2 = competency proceedings (experts-heavy), D3 = hearing notices (officer-heavy), F1 = psychiatric evaluation content (dsm5-heavy).

Connecting Context

Part 1 showed A-family documents are exclusively complaints; D3 is 88% hearing notices; C1 concentrates in Rule 20 orders. The string_group_type distributions match these document functions with precision. See Appendix A7.

Finding 7: Row-Level Text Overlap Between Groups

Background

While term-level analysis measures vocabulary, row-level text overlap (via `hashpack_rows` SHA256 hashes) measures whether documents in different groups contain identical text lines — a stronger signal of shared template content.

Findings

Row-level text uniqueness by group:

Group	Total Row Hashes	Distinct Hashes	Uniqueness Ratio	Template Reuse %	Rank
D2	10,784	753	7.0%	93.0%	1
C1	103,468	10,242	9.9%	90.1%	2
D3	14,918	1,726	11.6%	88.4%	3
D1	57,884	7,380	12.7%	87.3%	4
B1	37,653	7,069	18.8%	81.2%	5
A2	10,753	3,727	34.7%	65.3%	6
A3	7,802	2,768	35.5%	64.5%	7
A4	9,617	3,475	36.1%	63.9%	8
F1	2,350	868	36.9%	63.1%	9
A5	1,051	502	47.8%	52.2%	10
A1	889	462	52.0%	48.0%	11
E1	234	158	67.5%	32.5%	12
A6	184	158	85.9%	14.1%	13

D2 has the lowest uniqueness ratio at 7.0%: 93.0% of its 10,784 text lines are duplicated across its 131 documents. This is extreme template reuse — consistent with highly standardized Rule 20.01 evaluation orders. C1 (9.9%) and D1 (12.7%) show similar but less extreme patterns.

A6 has the highest uniqueness (85.9%), followed by E1 (67.5%) and A1 (52.0%). Smaller groups naturally have higher uniqueness ratios, but the magnitude of the difference (A6: 85.9% vs D2: 7.0%) reflects genuinely different template-reuse levels.

Top 10 cross-group shared row hash pairs (identical text lines appearing in both groups):

Group 1	Group 2	Shared Row Hashes	Boundary Type	Cross- Family?
C1	D1	1,681	Brick wall adjacent	Yes
A2	A4	1,288	Intra-family	No (intra-family)
B1	C1	1,019	Brick wall adjacent	Yes
C1	D3	997	Cross-family non-adjacent	Yes
A3	A4	800	Intra-family	No (intra-family)
D1	D2	691	Intra-family	No (intra-family)
B1	D1	646	Cross-family non-adjacent	Yes

C1	D2	643	Cross-family non-adjacent	Yes
A2	A3	418	Intra-family	No (intra-family)
A4	C1	398	Cross-family non-adjacent	Yes

Significance

The row-level sharing pattern mirrors the Jaccard term similarity: the strongest cross-group linkage is C1 ↔ D1 (1,681 shared lines), followed by B1 ↔ C1 (1,019) and B1 ↔ D1 (646). These groups share not just vocabulary but actual identical text — the same template paragraphs appearing verbatim across groups. The intra-family A2 ↔ A4 sharing (1,288 lines) exceeds most cross-family sharing, confirming strong within-family template reuse.

Connecting Context

Part 1 showed C1 and D1 share filing type concentrations in Rule 20 evaluation orders. The 1,681 shared text lines confirm these two groups produce documents from overlapping or related templates. The absence of strong row-level sharing at brick wall boundaries (A → B: 17–22 shared lines for A6/A1 pairs with B1) reinforces the structural isolation finding. See Appendix A8.

Finding 8: Document-Level Term Concentration (Pareto Analysis)

Background

If term occurrences are evenly distributed, each document in a group contributes proportionally to the term count. If they are concentrated, a small number of documents account for a disproportionate share — suggesting either dense evaluation reports or documents with unusual content.

Findings

The single highest term-density document per group:

Grp	Case ID	Filing Type	Filing Date	Term Rows	Dist Terms	Notes
A1	27-CR-17-22909	E-filed Comp-Warrant	2017-09-12	11	6	—
A2	27-CR-22-20527	Memorandum	2023-12-15	236	65	—
A3	27-CR-23-512	E-filed Comp-Order	2023-01-06	23	6	—
A4	27-CR-23-1886	Notice of Appeal	2025-05-29	248	83	Guertin case
A5	27-CR-23-13960	E-filed Comp-Order	2023-07-05	9	5	—
A6	27-CR-23-1101	Amended Complaint	2023-04-12	2	1	—
B1	27-CR-23-1886	Other Document	2025-04-28	5,561	89	Guertin case
C1	27-CR-23-1886	Other Document	2025-04-28	2,725	105	Guertin case
D1	27-CR-23-1886	Other Document	2025-04-28	2,250	36	Guertin case
D2	27-CR-22-18209	Rule 20.01 Eval Order	2022-10-13	59	21	—
D3	27-CR-21-23628	Proposed Order	2022-03-11	45	15	—
E1	27-CR-23-1886	Incompetency Finding	2024-01-17	89	34	Guertin case
F1	27-CR-23-1886	Other Document	2025-04-28	915	107	Guertin case

Case 27-CR-23-1886 (Guertin) produces the highest term-density document in 8 of 13 groups (B1, C1, D1, A4, E1, F1, and likely others). The B1 peak document alone contains 5,561 term occurrences across 89 distinct terms. This concentration is consistent with the “Other Document” filing type used for pro se filings with extensive legal argumentation and factual narratives.

Significance

The extreme concentration of peak term density in a single case across multiple groups indicates that case 27-CR-23-1886 generates documents with unusually rich extracted-term content. This does not invalidate group-level statistics (the per-group metrics are computed across all documents), but it identifies the case as a statistical outlier in term density.

Connecting Context

Part 1 identified case 27-CR-23-1886 as appearing across multiple tracking font groups. The term density analysis confirms that this case’s documents are not only structurally diverse (spanning

multiple font groups) but also terminologically dense — consistent with extensive pro se filings that contain detailed legal and factual content. See Appendix A9.

Finding 9: Group Language Signature Cards

Background

Analogous to Part 1's structural signature cards, these language fingerprints summarize each group's terminological identity in a standardized format.

Findings

A1: *Minimal-vocabulary complaint group; all terms are ubiquitous corpus-wide.*

Total term rows	36	Exclusive terms	0
Distinct terms	9	Exclusive % of vocab	0%
Vocab richness	0.2500	Ubiquitous term %	100%
Top category	itinerary (72.2%)	Row uniqueness	52.0%

A2: *Mid-size complaint group with modest exclusive vocabulary; fraud/surveillance terms distinguish it from other A-subgroups.*

Total term rows	775	Exclusive terms	6
Distinct terms	83	Exclusive % of vocab	7.2%
Vocab richness	0.1071	Ubiquitous term %	66.3%
Top category	itinerary (31.1%)	Row uniqueness	34.7%

A3: *Complaint group dominated by detention/itinerary language; one exclusive term (schizophrenic).*

Total term rows	310	Exclusive terms	1
Distinct terms	29	Exclusive % of vocab	3.4%
Vocab richness	0.0935	Ubiquitous term %	79.3%
Top category	itinerary (62.6%)	Row uniqueness	35.5%

A4: *Richest A-family vocabulary; 10 exclusive terms including forensic navigator and statutory references. Bridges toward B/C/D vocabulary.*

Total term rows	512	Exclusive terms	10
Distinct terms	104	Exclusive % of vocab	9.6%
Vocab richness	0.2031	Ubiquitous term %	63.5%
Top category	itinerary (23.8%)	Row uniqueness	36.1%

A5: *Small complaint group; all terms also found in other groups.*

Total term rows	39	Exclusive terms	0
Distinct terms	13	Exclusive % of vocab	0%
Vocab richness	0.3333	Ubiquitous term %	100%
Top category	itinerary (61.5%)	Row uniqueness	47.8%

A6: Single-document, single-term group. Maximum vocabulary poverty.

Total term rows	2	Exclusive terms	0
Distinct terms	1	Exclusive % of vocab	0%
Vocab richness	0.5000	Ubiquitous term %	100%
Top category	itinerary (100%)	Row uniqueness	85.9%

B1: Largest exclusive vocabulary (45 terms). Named examiners + forensic procedural terms. Broadest filing type coverage produces broadest vocabulary.

Total term rows	16,592	Exclusive terms	45
Distinct terms	219	Exclusive % of vocab	20.5%
Vocab richness	0.0132	Ubiquitous term %	38.8%
Top category	experts (25.4%)	Row uniqueness	18.8%

C1: Dominant competency pipeline group. Deep experts/protect/indefinite vocabulary. Named examiner exclusives (Amanda Powers, Jennifer Flowers).

Total term rows	45,227	Exclusive terms	18
Distinct terms	204	Exclusive % of vocab	8.8%
Vocab richness	0.0045	Ubiquitous term %	39.2%
Top category	experts (30.3%)	Row uniqueness	9.9%

D1: Evaluation-order group with highest experts concentration (47.1%). Exclusive terms are legal doctrine (metadata, Brady, informed consent) from pro se filings.

Total term rows	29,100	Exclusive terms	4
Distinct terms	143	Exclusive % of vocab	2.8%
Vocab richness	0.0049	Ubiquitous term %	54.5%
Top category	experts (47.1%)	Row uniqueness	12.7%

D2: Strict vocabulary subset of D1 (all 31 terms also in D1). Maximum template reuse. Most standardized document group.

Total term rows	6,319	Exclusive terms	0
Distinct terms	31	Exclusive % of vocab	0%
Vocab richness	0.0049	Ubiquitous term %	87.1%
Top category	experts (47.7%)	Row uniqueness	7.0%

D3: Linguistic outlier within D-family. Judicial officer names dominate; vocabulary inconsistent with sibling groups D1/D2.

Total term rows	369	Exclusive terms	1
Distinct terms	46	Exclusive % of vocab	2.2%
Vocab richness	0.1247	Ubiquitous term %	82.6%

Top category	officer (41.2%)	Row uniqueness	11.6%
---------------------	-----------------	-----------------------	-------

E1: *Micro-group (2 docs). Competency language subset. Zero exclusive terms — entirely embedded in the B/C/D vocabulary space.*

Total term rows	170	Exclusive terms	0
Distinct terms	38	Exclusive % of vocab	0%
Vocab richness	0.2235	Ubiquitous term %	84.2%
Top category	protect (27.6%)	Row uniqueness	67.5%

F1: *Highest exclusive density per document (4.0 terms/doc). DSM-5 diagnostic language dominates. Named examiners (Barbo, Bruss) exclusive. Clinical evaluation content from only 4 docs.*

Total term rows	1,104	Exclusive terms	16
Distinct terms	110	Exclusive % of vocab	14.5%
Vocab richness	0.0997	Ubiquitous term %	54.5%
Top category	dsm5 (21.8%)	Row uniqueness	36.9%

Finding 10: Letter-Family Language Synthesis

Background

Rolling up all 13 groups into 6 letter-families reveals the macro-level linguistic structure of the tracking font corpus.

Findings

Family-level Jaccard term similarity matrix:

	A	B	C	D	E	F
A	—	0.408	0.427	0.485	0.164	0.305
B	0.408	—	0.544	0.448	0.168	0.343
C	0.427	0.544	—	0.594	0.186	0.377
D	0.485	0.448	0.594	—	0.224	0.370
E	0.164	0.168	0.186	0.224	—	0.287
F	0.305	0.343	0.377	0.370	0.287	—

Family-level exclusive term counts and vocabulary sizes:

Family	Docs	Term Rows	Distinct Terms	Family-Excl Terms	Excl % of Vocab
A	139	1,674	147	19	12.9%
B	235	16,592	219	45	20.5%
C	1,089	45,227	204	18	8.8%
D	473	35,788	153	5	3.3%
E	2	170	38	0	0.0%
F	4	1,104	110	16	14.5%

The C ↔ D family pair is the most linguistically similar (Jaccard 0.594). B ↔ C is second (0.544). These three families (B, C, D) form a linguistic cluster centered on competency evaluation vocabulary. Family E is the most isolated — its Jaccard values with all other families range from 0.164 to 0.287, reflecting its status as a 2-document micro-group with no exclusive vocabulary.

Family B has the highest family-exclusive term count (45 terms, 20.5% of its vocabulary) despite being mid-sized by document count. Family F has the second-highest exclusive rate (14.5%) from only 4 documents. Family D has the lowest exclusive rate (3.3%) — its 153 terms are almost entirely shared with other families, particularly C.

Significance

The family-level synthesis confirms a three-tier linguistic architecture: (1) The B–C–D competency cluster with high mutual similarity and shared Rule 20 vocabulary; (2) Family A as the charging-instrument outlier with itinerary-dominated language; (3) Family F as a small but linguistically distinctive group with unique DSM-5 diagnostic content. Family E is too small to constitute a meaningful tier.

Connecting Context

Part 1's structural analysis identified the same three-tier pattern: B–C–D share judicial signatures and metadata pipelines; A-family uses a different producer pipeline (Winnovative); F1 has uniquely high signature density. The language analysis independently converges on the same structure. See Appendix A10.

LIMITATIONS & ALTERNATIVE EXPLANATIONS

Coverage gaps: 98 of 1,089 C1 documents (9.0%) and 32 of 238 D3 documents (13.4%) have zero `children_doc_strings` rows. These coverage gaps do not invalidate group-level statistics but may undercount C1 and D3 vocabulary sizes.

Term extraction method: The `children_doc_strings` table contains pre-extracted and classified terms, not raw full-text. Term exclusivity findings reflect this specific extraction pipeline and may not capture all vocabulary present in the original PDF text.

Small group sizes: Groups A1 (7 docs), A5 (6 docs), A6 (1 doc), E1 (2 docs), and F1 (4 docs) have very small sample sizes. Their zero exclusive term counts may reflect insufficient document volume rather than genuine vocabulary overlap.

Jaccard sensitivity to vocabulary size: Groups with very small vocabularies (A6: 1 term) will show near-zero Jaccard with all other groups regardless of actual semantic similarity. The A6 boundary findings should be interpreted with this limitation.

Benign template explanations: High template reuse (D2: 93% duplicate rows) is expected in standardized court order production. Identical text across groups may reflect shared court templates rather than any anomaly.

Examiner name exclusivity: Named examiners exclusive to specific groups (Sonia Reardon in B1; Amanda Powers in C1) may reflect normal assignment patterns — different examiners being active in different time periods or case types — rather than pipeline anomalies.

RECOMMENDATIONS FOR INDEPENDENT VERIFICATION

1. Re-extract term data independently from the original PDF files using a different text extraction pipeline (e.g., `pdftotext` + custom NLP) and compare term inventories per group against the `children_doc_strings` data.
2. Verify exclusive term findings by spot-checking the original PDF content for 5–10 exclusive terms per group to confirm they actually appear in the document text.
3. Extend the analysis to the full 4,251-document corpus to determine whether terms exclusive to the 13-group tracking font cohort also appear in non-tracked documents.
4. Conduct a temporal analysis of vocabulary shifts — do exclusive terms cluster in specific time periods, or are they distributed across the full filing date range?
5. Submit the Jaccard similarity matrices and language shift scores to a computational linguistics expert for independent validation of the boundary discontinuity findings.
6. All SQL queries used in this analysis are provided in the Appendix and can be executed against the live database by any party with READ access.

APPENDIX | REPRODUCIBLE SQL QUERIES

A1 — Corpus Scout — Cohort Census by PDF Group

```
SELECT pdf_group, font_group, COUNT(*) AS docs,
       COUNT(DISTINCT case_uid) AS cases,
       COUNT(DISTINCT cluster_id) AS clusters,
       MIN(filing_date) AS earliest,
       MAX(filing_date) AS latest
FROM children_tracking_groups
WHERE pdf_group IN ('A1', 'A2', 'A3', 'A4', 'A5', 'A6',
                   'B1', 'C1', 'D1', 'D2', 'D3', 'E1', 'F1')
GROUP BY pdf_group, font_group
ORDER BY pdf_group;
```

A2 — doc_strings Coverage by PDF Group

```
WITH grp AS (
  SELECT DISTINCT evidence_uid, pdf_group
  FROM children_tracking_groups
  WHERE pdf_group IN ('A1', 'A2', 'A3', 'A4', 'A5', 'A6',
                     'B1', 'C1', 'D1', 'D2', 'D3', 'E1', 'F1')
)
SELECT g.pdf_group,
       COUNT(DISTINCT g.evidence_uid) AS total_docs,
       COUNT(DISTINCT g.evidence_uid)
         FILTER (WHERE ds.evidence_uid IS NOT NULL)
         AS docs_with_strings,
       COUNT(DISTINCT ds.string_norm) AS dist_norms,
       COUNT(ds.children_doc_strings_row_id) AS total_rows
FROM grp g
LEFT JOIN children_doc_strings ds
  ON ds.evidence_uid = g.evidence_uid
GROUP BY g.pdf_group
ORDER BY g.pdf_group;
```

A3 — Vocabulary Inventory per Group

```
WITH grp AS (
  SELECT DISTINCT evidence_uid, pdf_group
  FROM children_tracking_groups
  WHERE pdf_group IN ('A1', 'A2', 'A3', 'A4', 'A5', 'A6',
                     'B1', 'C1', 'D1', 'D2', 'D3', 'E1', 'F1')
)
SELECT g.pdf_group,
       COUNT(DISTINCT ds.string_norm) AS dist_norms,
```

```

COUNT(DISTINCT ds.string_group_type) AS dist_sgt,
COUNT(*) AS total_rows,
ROUND(COUNT(DISTINCT ds.string_norm)::numeric
      / NULLIF(COUNT(*), 0), 4) AS vocab_richness
FROM children_doc_strings ds
JOIN grp g ON g.evidence_uid = ds.evidence_uid
GROUP BY g.pdf_group
ORDER BY dist_norms DESC;

```

A4 — Group-Exclusive Terms (Core CTE)

```

WITH grp AS (
  SELECT DISTINCT evidence_uid, pdf_group
  FROM children_tracking_groups
  WHERE pdf_group IN ('A1','A2','A3','A4','A5','A6',
    'B1','C1','D1','D2','D3','E1','F1')
),
term_group AS (
  SELECT ds.string_norm, ds.string_group_type,
    g.pdf_group,
    COUNT(*) AS occ_rows,
    COUNT(DISTINCT ds.evidence_uid) AS doc_count
  FROM children_doc_strings ds
  JOIN grp g ON g.evidence_uid = ds.evidence_uid
  GROUP BY ds.string_norm, ds.string_group_type,
    g.pdf_group
),
term_spread AS (
  SELECT string_norm, string_group_type,
    COUNT(DISTINCT pdf_group) AS n_groups,
    ARRAY_AGG(DISTINCT pdf_group ORDER BY pdf_group)
      AS groups_list,
    SUM(occ_rows) AS total_rows,
    SUM(doc_count) AS total_docs
  FROM term_group
  GROUP BY string_norm, string_group_type
)
SELECT groups_list[1] AS exclusive_group,
  string_norm, string_group_type,
  total_rows, total_docs
FROM term_spread
WHERE n_groups = 1
ORDER BY exclusive_group, total_rows DESC;

```

A5 — Pairwise Jaccard Similarity

```

WITH grp AS (
  SELECT DISTINCT evidence_uid, pdf_group
  FROM children_tracking_groups
  WHERE pdf_group IN ('A1','A2','A3','A4','A5','A6',
    'B1','C1','D1','D2','D3','E1','F1')

```

```

),
group_terms AS (
  SELECT DISTINCT g.pdf_group, ds.string_norm
  FROM children_doc_strings ds
  JOIN grp g ON g.evidence_uid = ds.evidence_uid
),
pair_intersect AS (
  SELECT a.pdf_group AS g1, b.pdf_group AS g2,
    COUNT(*) AS shared_terms
  FROM group_terms a
  JOIN group_terms b ON a.string_norm = b.string_norm
    AND a.pdf_group < b.pdf_group
  GROUP BY a.pdf_group, b.pdf_group
),
group_sizes AS (
  SELECT pdf_group, COUNT(*) AS term_count
  FROM group_terms GROUP BY pdf_group
)
SELECT pi.g1, pi.g2, pi.shared_terms,
  gs1.term_count + gs2.term_count - pi.shared_terms
  AS union_terms,
  ROUND(pi.shared_terms::numeric /
    (gs1.term_count + gs2.term_count
    - pi.shared_terms), 4) AS jaccard
FROM pair_intersect pi
JOIN group_sizes gs1 ON gs1.pdf_group = pi.g1
JOIN group_sizes gs2 ON gs2.pdf_group = pi.g2
ORDER BY jaccard DESC;

```

A6 — Brick Wall Language Shift Scores

```

-- See Finding 5 methodology.
-- Uses same group_terms CTE as A5, then computes
-- per-boundary exclusive-left, exclusive-right,
-- and shift score = (excl_left + excl_right) / union.

```

A7 — string_group_type Distribution by Group

```

WITH grp AS (
  SELECT DISTINCT evidence_uid, pdf_group
  FROM children_tracking_groups
  WHERE pdf_group IN ('A1', 'A2', 'A3', 'A4', 'A5', 'A6',
    'B1', 'C1', 'D1', 'D2', 'D3', 'E1', 'F1')
),
sgt_counts AS (
  SELECT g.pdf_group, ds.string_group_type,
    COUNT(*) AS cnt
  FROM children_doc_strings ds
  JOIN grp g ON g.evidence_uid = ds.evidence_uid
  GROUP BY g.pdf_group, ds.string_group_type
),

```

```

grp_totals AS (
  SELECT pdf_group, SUM(cnt) AS total
  FROM sgt_counts GROUP BY pdf_group
)
SELECT sc.pdf_group, sc.string_group_type, sc.cnt,
  ROUND(100.0 * sc.cnt / gt.total, 1) AS pct
FROM sgt_counts sc
JOIN grp_totals gt ON gt.pdf_group = sc.pdf_group
ORDER BY sc.pdf_group, sc.cnt DESC;

```

A8 — Cross-Group Row Hash Sharing

```

WITH grp AS (
  SELECT DISTINCT evidence_uid, pdf_group
  FROM children_tracking_groups
  WHERE pdf_group IN ('A1','A2','A3','A4','A5','A6',
    'B1','C1','D1','D2','D3','E1','F1')
),
hash_group AS (
  SELECT DISTINCT hr.row_sha256, g.pdf_group
  FROM hashpack_rows hr
  JOIN grp g ON g.evidence_uid = hr.evidence_uid
)
SELECT a.pdf_group AS g1, b.pdf_group AS g2,
  COUNT(*) AS shared_hashes
FROM hash_group a
JOIN hash_group b ON a.row_sha256 = b.row_sha256
  AND a.pdf_group < b.pdf_group
GROUP BY a.pdf_group, b.pdf_group
ORDER BY shared_hashes DESC;

```

A9 — Document-Level Term Concentration

```

WITH grp AS (
  SELECT DISTINCT evidence_uid, pdf_group
  FROM children_tracking_groups
  WHERE pdf_group IN ('A1','A2','A3','A4','A5','A6',
    'B1','C1','D1','D2','D3','E1','F1')
),
doc_density AS (
  SELECT g.pdf_group, ds.evidence_uid,
    ds.case_id, ds.filing_type, ds.filing_date,
    COUNT(*) AS term_rows,
    COUNT(DISTINCT ds.string_norm) AS dist_terms,
    ROW_NUMBER() OVER (PARTITION BY g.pdf_group
      ORDER BY COUNT(*) DESC) AS rn
  FROM children_doc_strings ds
  JOIN grp g ON g.evidence_uid = ds.evidence_uid
  GROUP BY g.pdf_group, ds.evidence_uid,
    ds.case_id, ds.filing_type, ds.filing_date
)

```

```
SELECT * FROM doc_density WHERE rn = 1  
ORDER BY pdf_group;
```

A10 — Letter-Family Jaccard Similarity

```
-- Same pattern as A5 but aggregated  
-- at font_group / family level  
-- (A=TTF-A, B=TTF-B, C=TTF-C, D=TTF-D,  
--  E=TTF-E, F=TTF-F)
```

Source: *Matthew Guertin v. State of Minnesota* | 27-CR-23-1886 | MCRO Forensic Database | MattGuertin.com