Information-Theoretic Identifiability of Split-Mark Diagnostic Posets, L1 Micro-Syntax Transfer Independence, and Sentence-Level Propositional Conservation in English Prose (1818-2026)

Gareth Montgomery

Sprucefield Research & GarethEnglish.com, Lisburn, Northern Ireland, UK · October 2026

Abstract. This monograph establishes four interconnected results across cognitive diagnostic psychometrics, second-language acquisition (SLA) corpus linguistics, and quantitative diachronic philology. First, in Cambridge C1/C2 Key Word Transformation (Part 4) testing, each item awards up to 2 marks across two syntactic loci, yet standard reporting collapses the ordered binary vector $(m_L, m_R) \in \{0,1\}^2$ into the scalar sum $S = m_L + m_R \in \{0,1,2\}$. Over a 5-attribute hierarchical syntactic poset with $|S| = 10$ permissible mastery states (prior Shannon entropy $\log_2 10 \approx 3.3219$ bits), we prove that ordered Split-Mark vector scoring achieves full poset identifiability in $N_{min}(V) = 3$ static items (and 2.3000 expected items adaptively), compared with $N_{min}(S) = 4$ for scalar sums and $N_{min}(D) = 5$ for dichotomous scoring. Second, auditing $N = 80$ Cambridge C2 Proficiency Part 4 items and 14 longitudinal C1/C2 scripts (3,886 words across 322 sentences) from zero-article Slavic/Baltic (Russian, Ukrainian, Lithuanian) and Romance (Italian, French) L1 learners reveals the Macro/Micro Decoupling Paradox: candidates attain 96.9% accuracy on C1 macro-syntactic frames while remaining capped at a 21.4% Error-Free Sentence Ratio (EFSR) by deterministic L1 micro-syntax transfer (50.8% determiner errors in Slavic/Baltic vs 26.8% in Romance), while macro-frame deployment is statistically invariant to L1 family ($\chi^2 = 0.3113$, $p = 0.7258$). Third, we introduce a closed-class stylometric discriminator combining Clause-Initial Opener Shannon Entropy (H_{opener}) and Punctuation Markov Transition Entropy (H_{punct}) that separates authentic C1/C2 prose from formulaic template prose with $AUC = 1.000$ (Mann-Whitney $p = 1.08 \times 10^{-5}$). Fourth, across a balanced 36-text diachronic corpus (5,296 words, 1800-2026), we prove Sentence-Level Propositional Conservation: while finite subordination density decayed by 52.6% ($t = -5.54$, $p = 1.98 \times 10^{-5}$), nominal-participial compression surged by 240.0% ($t = +8.86$, $p = 1.43 \times 10^{-8}$), holding total propositions per sentence invariant across two centuries (2.70 → 2.92, $p = 0.4778$).

1. Split-Mark Poset Identifiability in Cognitive Diagnostic Models

Consider a diagnostic assessment whose attribute space $A = \{a_1, \ldots, a_K\}$ is governed by a prerequisite poset $(A, \preceq)$, restricting the state space to the lattice of order ideals $S = J(A) \subseteq \{0,1\}^K$. In Cambridge C1 Advanced and C2 Proficiency Use of English Part 4, each item consists of two syntactic sub-tasks (left locus L_j and right locus R_j) with conjunctive DINA responses $m_{j,L}(\alpha) = \prod_{k \in q_{j,L}} \alpha_k$ and $m_{j,R}(\alpha) = \prod_{k \in q_{j,R}} \alpha_k$.

> **Theorem 1.1 (Split-Mark Information Dominance & Minimum Test Length).** Let Π_D, Π_S, Π_V denote the state partitions of S induced by dichotomous scoring $D_j = m_{j,L}\, m_{j,R}$, scalar sum scoring $S_j = m_{j,L} + m_{j,R}$, and ordered Split-Mark vector scoring $V_j = (m_{j,L}, m_{j,R})$. Then Π_V strictly refines Π_S, which refines Π_D, so that $I(\Theta; V) \geq I(\Theta; S) \geq I(\Theta; D)$. Over the 5-attribute Cambridge

> syntactic poset with $|S| = 10$ permissible states, the minimum test lengths for complete state identifiability ($I = \log_2 10 \approx 3.3219$ bits) are:
>
> $$N_{min}(V) = 3 \text{ items}, \quad N_{min}(S) = 4 \text{ items}, \quad N_{min}(D) = 5 \text{ items}, \quad E[N_{adaptive}(V)] = 2.3000 \text{ items}.$$

Scoring Regime	3-Item Equivalence Classes	Singletons Identified	Shannon Information (bits)	Min Items for 10/10 States	Adaptive Expected Items
Dichotomous (0/1)	6 / 10	2 / 10	2.5219 bits (75.9%)	5 items	3.7000 items
Scalar Sum (0/1/2)	9 / 10	8 / 10	3.1219 bits (94.0%)	4 items	2.5000 items
Ordered Split-Mark Vector (m_L, m_R)	**10 / 10**	**10 / 10**	**3.3219 bits (100.0%)**	**3 items**	**2.3000 items**

2. Empirical Audit of CPE Part 4 & The Macro/Micro Decoupling Paradox

An exhaustive 2-locus syntactic audit of $N = 80$ Cambridge C2 Proficiency Part 4 transformations (160 mark-bearing loci across the 50-item checklist and 30-item diagnostic battery) demonstrates that **80.0%** (64/80) of items test verb-to-noun nominalisation, fixed C2 idiom/collocation frames, or dependent prepositions as their primary target, and **85.0%** (68/80) require an obligatory dependent preposition in at least one mark locus.

> **Theorem 2.1 (Macro/Micro Transfer Independence).** Across 14 longitudinal C1/C2 essays (3,886 words, 322 sentences), let $M(\alpha)$ denote C1 macro-syntactic frame accuracy (clefts, negative inversions, passive reporting, participle clauses) and $E(\mu \mid L_1)$ denote micro-syntactic accuracy (articles, prepositions, morphosyntactic agreement). Then:
>
> $$P(M = 1) = 96.9\% \text{ (31/32 frames)}, \quad \chi^2(M \perp L_1) = 0.3113 \ (p = 0.7258), \quad EFSR_{raw} = 21.4\% \rightarrow EFSR_{scrubbed} = 57.1\%.$$

L1 Typological Group	Scripts (Words / Sents)	C1 Macro-Frame Accuracy	Determiner / Article Share	Pre-Scrub EFSR	Post-L1-Scrub EFSR
Slavic / Baltic (RU, UK, LT)	8 (2,163w / 181s)	94.7% (18 / 19)	**50.8%** (30 / 59 errors)	18.8%	55.8% (+37.0 pp)
Romance (IT, FR)	6 (1,723w / 141s)	100.0% (13 / 13)	**26.8%** (15 / 56 errors)	24.8%	58.9% (+34.1 pp)
Combined Longitudinal Corpus	**14 (3,886w / 322s)**	**96.9% (31 / 32)**	**39.1% (45 / 115 errors)**	**21.4%**	**57.1% (+35.7 pp)**

3. Closed-Class Syntactic Cadence Entropy vs Template Over-Stuffing

We define the **Syntactic Cadence Quality Score (SCQS)** strictly from closed-class functional distributions (clause-initial opener entropy H_{opener} over 8 syntactic categories, punctuation Markov transition entropy H_{punct}, inter-sentence length dispersion CV_{len}, and macro-shift density penalty around the optimal band *[0.18, 0.38]* shifts per 100 words):

$$\text{SCQS} = \text{clip}_{[0,100]}(22\ H_{\text{opener}} + 20\ H_{\text{punct}} + 38\ \min(\text{CV}_{\text{len}} / 0.42, 1.15) - P_{\text{stuff}}(\rho_{\text{shift}}) - P_{\text{canned}}).$$

Across 28 benchmark texts (14 authentic C1/C2 scripts vs 14 template/memorised-connector essays), traditional lexical diversity (MATTR-50) fails to separate authentic from formulaic writing (*0.793 vs 0.808, $p = 0.0847$*), whereas **SCQS achieves complete separation (mean 76.22 vs 18.88, Cohen's d = 7.72, ROC AUC = 1.000, Mann-Whitney $p = 1.08 \times 10^{-5}$)**.

4. Diachronic Sentence-Level Propositional Conservation (1818-2026)

> **Theorem 4.1 (Propositional Conservation under Syntactic Compression).** Across a 36-text balanced diachronic corpus (5,296 words across Period I: 1800-1899, Period II: 1900-1979, and Period III: 1990-2026), let *F(s)* denote finite subordinate clauses per sentence and *C(s)* denote non-finite participial, gerund-prepositional, and deverbal nominalisation phrases per sentence. While *F(s)* decayed by *52.6%* and *C(s)* rose by *240.0%*, the total propositional load *PPS(s) = 1 + F(s) + C(s)* is statistically conserved:
>
> $$\text{PPS}_{1800\text{-}1899} = 2.70 \pm 0.34 \rightarrow \text{PPS}_{1900\text{-}1979} = 2.77 \pm 0.28 \rightarrow \text{PPS}_{1990\text{-}2026} = 2.92 \pm 0.25$$
> $$(\text{ANOVA } F = 0.76,\ p = 0.4778).$$

Diachronic Stratum	Mean Sentence Length	Finite Subordination / Sent	Nominal-Participial / Sent	Compression Ratio (NP / Sub)	Propositions / Sent (PPS)
Period I: Late Georgian / Victorian (1800-1899)	36.33 words	1.364	0.333	0.244	**2.697**
Period II: Early-Mid 20th Century (1900-1979)	28.20 words	0.927	0.841	0.907	**2.768**
Period III: Contemporary C1/C2 (1990-2026)	23.94 words (-34.1%)	0.646 (-52.6%, p=1.98e-5)	1.133 (+240.0%, p=1.43e-8)	1.754 (7.2×)	**2.921 (p=0.4778)**

References & Reproducibility

Full LaTeX source (`CL_Manuscript_Montgomery.tex`), interactive browser verifiers (`sprucefield-research.html` and `c1-syntactic-research-lab.html`), and deterministic corpus verification scripts are publicly archived at https://garethenglish.com/sprucefield-research.html.