レジスタ比率復元法

江原　遥
東京学芸大学
ehara@u-gakugei.ac.jp

概要

語彙学習では，学習者が実際に遭遇する言語環境に即した語の習得が重要である。一般コーパスは多様なレジスタを含むが，その比率は必ずしも学習向けに最適化されていない。一方，言語教育では人手判断などを用いたレベル別語彙リストが作成されてきたが，それが補うレジスタの偏りは明確でなかった。本研究は，一般コーパスと語彙リストから，語の頻度順が最も一致するレジスタ比率を線形計画法で推定する手法を提案する。英語・日本語で評価し，LLM 向け語彙リストの分析にも有効であることを示した。

1　はじめに

言語教育では，第一言語・第二言語学習者が実際に遭遇する言語環境に語彙学習を整合させることが長らく重視されてきた．教師や教材作成者（語彙リスト作成者）は，特徴的な「レジスタ」への接触が教室から実使用への移行（transfer）を促すと想定し，想定されるコミュニケーション領域に合わせて語彙を段階化したカリキュラムを設計する．レジスタとは，書き言葉・話し言葉のような言語使用の在り方を捉える言語学的概念である．

British National Corpus（BNC）のような一般コーパスは，多様な領域にわたる書き言葉・話し言葉資料を収集し，語彙頻度表を提供することで，この整合を支える基盤として教育・研究で広く再利用されてきた．これらのコーパスは異なるコミュニケーション状況をレジスタラベルとして束ねるが，コーパス内のレジスタ比率が学習者の日常的な接触環境の混合比率を反映するとは限らない．それにもかかわらず，頻度表，語彙プロフィール，自動難易度推定はしばしばコーパス全体の出現数から直接導出されるため，大規模な文書集合の影響で書き言葉レジスタが過大評価されやすい．この偏りを補正するために，実務者は独自の語彙リストを作成したり，学習者テストの結果に依拠したりするが，複数のシグナルをどのように整合させるべきかは必ずしも明示されてこなかった．

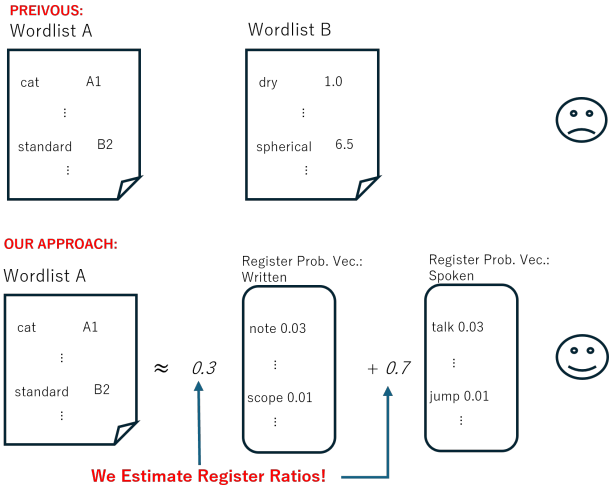

図 1　本研究の概要．従来は，教育用語彙リストごとに異なる指標が用いられ，想定する学習者像を理解するには作成者のマニュアルを参照する必要があった．その結果，学習者は学習の進度に応じてどの語彙リストを用いるべきか判断しにくかった．本研究は，レベル別語彙リストと一般コーパスが与えられたとき，語彙リストと最も整合するようなレジスタごとの頻度分布混合比率を推定する．これにより，学習者・教育者は学習に適した語彙リストを解釈しやすくなる．

混合比率推定は，不透明な埋め込み表現ではなく解釈可能なレジスタ重みベクトルを直接得られるため，教育的分析にとって魅力的である．本研究の重要な動機は，既存の語彙難易度資源が相互に比較不可能な尺度で難易度を符号化している点にある．段階的な熟達度尺度に整合したシラバスは熟練教師の判断を反映する一方，語彙複雑性に関する共有タスクや語彙サイズ評価は，連続値スコア，多数決，段階評定などにより学習者の知覚を捉える 文献 [1, 2, 3, 4]．提案枠組みが必要とするのは「語彙項目 w_i が w_j より易しい」といった対比較の順序情報のみであり，各資源の絶対的なキャリブレーションには依存しない．この尺度不変性により，教師作成の

シラバスと学習者由来の注釈を，同一の推定手続きにおいて再正規化や線形関係の仮定なしに統合できる．

図 1 は本研究の概観を示す．従来手法では，学習用語彙リストが与えられても，想定されている言語環境を把握することが難しい．さらに，語彙は語彙リストごとに異なる測定水準で記述されてきた．本研究は，語彙リストの背後にあるレジスタ比率を推定する手法を提案する．従来法と異なり，入力語彙リストがレベル分類されていればよく，分類は実数値でも段階値でもよい．

本研究の主な貢献は以下の四点である．

- 作成済みの語彙リストと一般コーパスが与えられたときに，尺度に依存せずレジスタ比率を推定する手法を提案する．
- 英語・日本語の一般コーパスにおいて，粗い（話し言葉／書き言葉）および細粒度のレジスタ設定の下で，専門家が整備した資源と学習者由来資源を広範に比較する実証研究を行い，同一パイプラインで大規模言語モデルが生成した語彙リストも監査できることを示す．公開物には，レジスタ別出現数のキャッシュ化スクリプトや，語彙サイズテストから推定した項目反応理論（IRT）推定値の統合スクリプトを含める．
- 尺度不変であるため，ChatGPT のような言語学習向け LLM に対して，特定状況（例：留学）を想定したプロンプトにより作成した語彙リストについても，レジスタ推定により評価できる．
- 提案手法は線形計画問題であるため，大規模設定でも計算可能であり，多様な語彙リストと一般コーパスへ適用できる．さらに，得られた最適解の質を語彙リストやコーパスに依存しない形で解釈できる．本研究で算出したレジスタ比率はデータセットとして公開する予定である．

これらの知見は，難易度の段階付けと観測されたレジスタ分布を整合させた読解教材の設計に資する．

2 提案手法

本研究のパイプラインは，語彙難易度資源をレジスタ混合を制約する不等式へ変換する．レジスタ集合を R，コーパスに観測され，かつ難易度資源にも含まれる語彙素（lexeme）の集合を V とする．BNC からレジスタ別出現数を計算し，粗い設定では $R = \{\text{spoken}, \text{written}\}$，細粒度設定では *written:fiction* や *spoken:convrsn* など 8 種類のレジスタを用いる．語彙素 w のレジスタ r における出現数を $c_r(w)$，レジスタ r の総トークン数を $C_r = \sum_w c_r(w)$ とする．条件付き確率は正規化出現数により $p(w \mid r) = c_r(w)/C_r$ と近似する．混合重み α_r は $\alpha_r \geq 0$ かつ $\sum_r \alpha_r = 1$ を満たす．混合下で w が観測される期待確率は $q(w) = \sum_r \alpha_r p(w \mid r)$ である．

2.1 制約生成

CEFR-J のように順序尺度の難易度レベルを与える資源は，「w_i が w_j より易しい」ことを表す順序付きペア (w_i, w_j) へ変換する．難易度が実数で与えられる資源（例：CompLex）については，ペアをサンプリングして順序制約へ変換する．制約は $q(w_i) + s_{ij} \geq q(w_j)$ とし，$s_{ij} \geq 0$ を「スラック変数」と呼ぶ．$s_{ij} = 0$ は制約が満たされている（すなわち，現在の α_r により語頻度の比較 (i, j) が語彙リストの順序に従う）ことを表す．一方，$s_{ij} > 0$ なら制約違反であり，値が大きいほど違反が大きい．このとき線形不等式は

$$\sum_{r \in R} \alpha_r p(w_i \mid r) \geq \sum_{r \in R} \alpha_r p(w_j \mid r) + s_{ij} \qquad (1)$$

となる．数値スコアを与える資源では，易しい語の確率がより大きくなるよう制約を対称化して加える．順序レベル資源は非対称のまま扱う．

2.2 最適化目的関数

s_{ij} が大きいほど「i が j より易しい」という順序制約の違反が大きいことに注意する．補助変数 t を導入し，次を解く：

$$\min_{\alpha, s, t} \; t \quad \text{s.t.} \quad s_{ij} \leq t, \; \forall (i, j). \qquad (2)$$

目的変数 t は最大違反の上界であり，制約数が 20 万ペアに増加しても混合比率を安定化する．また，最大スラックに到達する語ペアを特定できるため，誤差分析（どの対比較が矛盾の原因か）のために有用である．本論文では，メモリ 256 GB の計算機上で Gurobi を用いて最適化問題を解いた．ただし本問題は線形計画であるため，非商用ソルバ（例：`pulp`）でも解くことができる．

3 実験

すべての実験で，難易度指標から 1,000〜500,000 のの範囲でランダムサンプリングを用いて制約を作成し，スラック総和最小化解を求めた．

データセット　書き言葉・話し言葉のバランスを取って収集され，特定ジャンルまでのレジスタ注釈を含む British National Corpus（BNC）を用いた [5]．粗粒度実験ではレジスタを書き言葉と話し言葉の 2 レジスタに集約し，その比率を計算する．細粒度実験では，BNC の 8 レジスタ（*spoken:convrsn*，*spoken:othersp*，*written:acprose*，*written:fiction*，*written:news*，*written:nonac*，*written:otherpub*，*written:unpub*）の比率を計算する．さらにコーパス間の頑健性を検証するため，Open American National Corpus（OANC）[6] でも同一手順で実験した．OANC の話し言葉は *face-to-face* と *telephone* 会話から構成され，書き言葉は *fiction*，*journal*，*letters*，*non-fiction*，*technical*，*travel_guides* を含む．この二重設定により，コーパス固有のレジスタ目録が推定混合に実質的影響を与えるかを評価できる．

語彙難易度資源は次を用いた．CEFR-J は日本人英語学習者向けの見出し語に CEFR レベルを付与し，品詞別の変種も含む 文献 [7]．CompLex はクラウドワーカーの対比較を集約し，ある語彙素が別の語彙素より複雑である確率を近似する連続値スコアを与える 文献 [1]．CWI 2016 共有タスクは文レベル注釈を提供し，注釈者投票を集約して単語（ユニグラム）スコアへ変換できる 文献 [2]．CWI 2018 は Wikipedia ／ News ／ WikiNews の 3 分割を学習者が注釈したデータによりカバレッジを拡張した 文献 [3]．最後に，EVKD 語彙サイズテストは 100 名の被験者が 100 問の多肢選択に回答したデータであり，項目反応理論により項目パラメータを推定して連続的難易度を得る．

これらの資源は出自が大きく異なる．CEFR-J [7] は専門家教師によってカリキュラム期待に整合するよう整備された一方，CompLex [1]，CWI16 [2]，CWI18 [3]，EVKD [8] は学習者の判断や成績を捉える．その結果，尺度は直接比較できない：CEFR-J は離散的 CEFR 帯，CompLex は $[0, 1]$ の確率，CWI は整数投票の平均，EVKD は IRT 由来のロジットである．提案最適化は各信号を順序付きペアへ変換して一様に扱うことで，教師志向のシラバスと学習者中心の証拠を単一の混合推定問題で比較可能にする．

4　実験

各難易度資源から順序付き語彙ペア制約をランダム抽出（1,000–500,000）し，BNC のレジスタ別頻度に基づき目的 (2) の違反最小化で混合 α を推定した．

資源	非ゼロのレジスタ重み（5 万ペア）
CEFR-J	spoken:othersp 0.194, spoken:convrsn 0.024, written:fiction 0.609, written:news 0.136, written:unpub 0.037
CompLex	spoken:convrsn 0.170, written:acprose 0.202, written:fiction 0.628
CWI18 Wikipedia	spoken:othersp 0.102, written:acprose 0.382, written:fiction 0.336, written:otherpub 0.181
CWI18 News	spoken:convrsn 0.521, written:acprose 0.185, written:fiction 0.053, written:nonac 0.083, written:otherpub 0.159
CWI18 WikiNews	spoken:convrsn 0.112, written:fiction 0.888
VST 1PL	written:acprose 0.797, written:news 0.203
CWI16（共有語彙）	written:acprose 1.000

表 1　BNC 細粒度混合（目的=max，制約 5 万，seed 99）．

データセット　British National Corpus (BNC) を用い，(i) 書き言葉／話し言葉の粗粒度と，(ii) 8 レジスタの細粒度で推定した [5]．頑健性確認として同手順を OANC でも実施した [6]（レジスタ目録等は Appendix A）．

難易度資源　CEFR-J 文献 [7]，CompLex 文献 [1]，CWI16/18 文献 [2, 3]，EVKD 語彙サイズテスト 文献 [8] を用いた．尺度が異なるため，すべて順序付きペアへ変換して統一的に扱う（詳細は Appendix A）．

実験設定　各資源語彙を BNC 補題集合と交差し，一意ペアが尽きるまで制約を抽出した．レジスタ確率は補題出現数（スムージングなし）から計算し，制約は一様重みとした．主要結果は制約 5 万（seed 99）を示し，予算・感度・対照実験は Appendix A に回す．

細粒度比率復元結果　表 1 に 5 万制約の非ゼロ混合を示す．CEFR-J はフィクション優勢だが話し言葉質量を残し，CompLex は書き言葉側へ集中する．CWI18 は分割で傾向が異なり，News は会話レジスタを強く強調する一方，Wikipedia/WikiNews は書き言葉寄りである．共有語彙制約（CWI16）は学術散文へ退化した．

粗い混合　表 2 に 20 万制約の粗粒度混合を示す．CEFR-J は相対的に話し言葉を保持する一方，多くの学習者由来資源は書き言葉へ集中する．CWI18 News は話し言葉へ，IRT 由来推定は書き言葉へ崩壊した（追加診断は Appendix A）．

資源	話し言葉	書き言葉
CEFR-J	0.24	0.76
CompLex	0.19	0.81
CWI18 Wikipedia	0.19	0.81
CWI18 News	1.00	0.00
CWI18 WikiNews	0.15	0.85
CWI16	0.00	1.00
VST 1PL	0.00	1.00
VST 2PL Difficulty	0.00	1.00
VST 2PL Discrimination	0.00	1.00

表 2　BNC 粗粒度混合（目的=max，制約 20 万，seed 99）．

レジスタ	説明	Sakamoto	ShinSakamoto
OP	Opinion magazines	0.113	1.000
OL	Local newspapers	0.166	0.000
OM	General magazines	0.169	0.000
OY	Youth magazines	0.304	0.000
PN	National newspapers	0.247	0.000

表 3　BCCWJ 細粒度混合（目的=max，制約 5 万，seed 99）．

日本語におけるレジスタ復元　BCCWJ のレジスタ別頻度表と NINJAL 教育基本語彙（坂本／新坂本）を用い，英語と同様に順序付きペア制約から混合を推定した [9]．表 3 の通り，両リストとも推定混合に話し言葉レジスタが現れず，強い書き言葉偏りが観測された．また坂本は複数レジスタへ分散する一方，新坂本は単一レジスタへ収束した（詳細は Appendix A）．

LLM 作成語彙リストの評価　応用例として，同一人物像（計算機科学を学ぶ日本人学部生）に対し，(JP) 日本の大学院で英語論文を読む／ (US) 米国大学院で口頭コミュニケーション，の 2 条件で LLM に「5 段階× 100 語」の語彙リストを生成させ，レジスタ混合を推定した（プロンプト全文は Appendix A）．表 4 より，GPT-5 Pro は会話優勢だが条件により書き言葉側の内訳が変化し，Claude Sonnet は両条件で会話へ収束した．本結果は，自動生成教材を追加調整なしに監査できることを示す．

5　考察

実験は，語彙難易度資源が想定するレジスタ接触が一様でないことを示した．CEFR-J は制約が少ないと会話寄りだが，ペア数を増やすと書き言葉側へ移るため，小規模サブセットのみの参照は，上級学習者に必要な書き言葉接触を過小評価しうる．

同じ診断は自動生成教材にも適用でき（表 4），GPT-5 Pro は条件により書き言葉成分（学術／非学術）が入れ替わる一方，Claude Sonnet は会話へ収束した．

本手法は尺度不変で，教師作成（CEFR-J）と学習者由来（CompLex，CWI 等）を同一の対比較制約として扱い，不一致を復元混合として可視化する．また IRT で校正された語彙テストは受験者集団の影響を受けにくく 文献 [10, 4]，表 1 でも VST は学術散文／ニュースへ一貫して整合した．

学習者由来資源は概ね書き言葉優勢だが，CWI18 News のみ話し言葉へ崩壊するため，CWI18 分割の単純プールは推定を口頭側へ偏らせうる．本枠組みは LDA 等の教師なしトピック発見ではなく，既知レジスタで「難易度順序を最もよく説明する混合」を直接返す．

実装上，Lidstone スムージング（$\lambda \in \{0, 0.1, 1\}$）の有無で粗粒度混合は最大 2 ポイントしか変わらず，総和最小化はサンプルサイズ依存と退化が解消しないため，最大スラック最小化を採用した（例：5,000 語で約 1.2×10^7 ペア）．従って，本研究の定性的差は偶然的ゼロ頻度ではなく難易度証拠の不一致に起因すると考えられる．

6　結論

本研究では，語彙難易度制約からレジスタ混合を推定する手法を提案し，学習者志向資源の広範な集合に適用した．一般コーパスとして英語・日本語の 2 言語を用いて評価し，さらに LLM が生成した語彙リストに内在するレジスタ偏りを監査できることを示した．これにより，実務者は自動生成教材を配布前に迅速に点検できる．

今後の課題として，本手法は学習者目標に応じてレジスタ接触を調整する適応的読解プラットフォームを支援しうる．領域特化コーパスと対象集団から得た難易度判断を組み合わせることで，教材が学習者の想定するレジスタを反映しているかを迅速に評価できる．

謝辞

本研究は，科学技術振興機構さきがけ研究費 (JPMJPR2363)，JSPS 科研費 22K12287 の支援を受けた．

参考文献

[1] Matthew Shardlow, Marcos Zampieri, Mihael Pasov, and Cassandre Boulc. Semeval-2021 task 1: Lexical complexity prediction. In **Proceedings of the 15th International Workshop on Semantic Evaluation (SemEval-2021)**, pp. 1–16, Online, 2021. Association for Computational Linguistics.

[2] Gustavo Paetzold and Lucia Specia. Semeval 2016 task 11: Complex word identification. In **Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)**, pp. 560–569, San Diego, California, 2016. Association for Computational Linguistics.

[3] Seid Muhie Yimam, Chris Biemann, Gustavo Paetzold, and Lucia Specia. Multilingual and cross-lingual complex word identification. In **Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies**, pp. 401–411, New Orleans, Louisiana, 2018. Association for Computational Linguistics.

[4] Laura Palacios, Koji Yamamoto, and Yuki Sato. Evkd vocabulary size test response set. Educational Vocabulary Knowledge Diagnostic Project Dataset, 2018. Version 1.0.

[5] British National Corpus Consortium. British national corpus, version 3 (BNC XML edition), 2007. PID http://hdl.handle.net/20.500.12024/2554.

[6] Nancy Ide. The american national corpus: Then, now, and tomorrow. In **Selected Proceedings of the 2008 HCSNet Workshop on Designing the Australian National Corpus: Mustering Languages, Summerville, MA. Cascadilla Proceedings Project**, Vol. 127, 2008.

[7] Yukinori Tono, Satoko Kawaguchi, and Masashi Negishi. Developing a CEFR-based word list for Japanese learners. In **Research and Practice in Assessing Academic Writing**, pp. 55–76. Cambridge University Press, 2013.

[8] Yo Ehara. Building an english vocabulary knowledge dataset of japanese english-as-a-second-language learners using crowdsourcing. In **Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)**, 2018.

[9] Mai Omura and Masayuki Asahara. Ud-japanese bccwj: Universal dependencies annotation for the balanced corpus of contemporary written japanese. In **Proceedings of the Second Workshop on Universal Dependencies (UDW 2018)**, pp. 117–125, 2018.

[10] Chung-Hua Chen and Shu-Yu Li. Applying item response theory to vocabulary assessment. **Language Testing**, Vol. 28, No. 4, pp. 681–703, 2011.

[11] Ichiro Sakamoto. **Kyoiku Kihon Goi (Educational Basic Wordlists)**. Maki-Shoten, 1958.

[12] Ichiro Sakamoto. **Shin Kyoiku Kihon Goi (New Educational Basic Wordlists)**. Gakugei-Tosho, 1984.

A　実験詳細等

コーパスとレジスタ　BNC [5]：粗粒度は spoken/written，細粒度は 8 レジスタ（*spoken:convrsn, spoken:othersp, written:acprose, written:fiction, written:news, written:nonac, written:otherpub, written:unpub*）．頑健性として OANC [6] でも同手順（spoken=*face-to-face/telephone*；written=*fiction, journal, letters, non-fiction, technical, travel_guides*）．

難易度資源と尺度差　CEFR-J 文献 [7]，CompLex 文献 [1]，CWI16/18 文献 [2, 3]，EVKD VST 文献 [8] を使用．尺度（CEFR 帯／ [0, 1] ／投票平均／ IRT ロジット）が異なるため，絶対値は比較せず順序のみを対比較制約へ写像した．

追加設定・感度　語彙 ∩BNC からランダムに一意ペア制約を抽出（CEFR-J:1k–500k, CompLex/CWI:≤200k，VST: 数千で飽和）．レジスタ確率は補題頻度（無スムージング）で算出し，制約は一様重み．共有語彙条件は単一レジスタへ退化しうる（例：CWI16）．最小頻度閾値 1–5 でも混合は最大 2 ポイント変化に留まり，総和最小化は退化が解消しないため違反最小化を採用した．

追加結果・診断　VST 2PL は混合自体は 1PL と同様だが，識別力は学術散文–ニュース対比を，難易度は稀語彙を強調する．予算軌跡も大局不変（例：CEFR-J は会話優勢 → 書き言葉優勢へ移行）．目的値は CompLex/CWI で $< 10^{-4}$，CEFR-J で約 0.10 に留まり，最大違反は *onomatopoeia* 等の稀学術語に由来した．

日本語実験　BCCWJ の 13 レジスタ頻度を用い [9]，NINJAL 教育基本語彙（坂本／新坂本）を順序（A1 易–C4 難）へ写像して対比較制約化した．BCCWJ 非依存な旧版として [11, 12] を採用し，推定混合差を教育語彙の変化として比較した．

LLM プロンプト

- (JP) 日本の大学院進学・国際会議投稿に向け英語論文を読む日本人 CS 学部生向けに，英単語 100 語を易 → 難の 5 段階で作成せよ．
- (US) 米国大学院進学・母語話者と口頭コミュニケーションが必要な日本人 CS 学部生向けに，同様に作成せよ．

感度分析　サンプリング方式（random/prefix）と乱数シードを変えても資源間の相対順位は概ね不変で，seed 101 での成分変動は最大 0.02 に留まった．一方，prefix では易しい部分が強調され混合が極端化するが，違反最小化目的は常に最大スラックをより小さくし，監査で高影響の矛盾集合を明確化する．

コーパス	語彙リスト	語ペア	最大スラック
BNC	CEFR-J	*zoo* ≻ *yearn*	0.100
BNC	CompLex	*zone* ≻ *zither*	1.6×10^{-3}
BNC	CWI18 Wikipedia	*zoological* ≻ *vandalism*	3.7×10^{-4}
OANC	CEFR-J	*zoo* ≻ *widen*	0.100
OANC	CompLex	*zone* ≻ *vertebrates*	1.0×10^{-3}
OANC	CWI18 Wikipedia	*zoological* ≻ *volunteered*	2.4×10^{-4}

表 5　違反最小化目的における代表的実行の最大スラック．小さいほど，難易度資源とコーパスレジスタがよく整合している．

資源	Face-to-face	Fiction	Journal	Letters	Non-fiction	Technical	Telephone	Travel guides
CEFR-J	0.00	0.35	0.00	0.00	0.00	0.13	0.45	0.07
CompLex	0.20	0.21	0.00	0.00	0.00	0.34	0.00	0.25
CWI18 Wikipedia	0.04	0.46	0.32	0.03	0.00	0.02	0.00	0.13

表 6　違反最小化目的，ランダム制約 10 万（seed 99）で推定した OANC 混合．色の濃さは値の大きさを表す．

スラック診断　代表的実行の最大スラックを表 5 に示す．CEFR-J は 0.10 上界に到達するペアがあり，CompLex は 10^{-3} 程度，CWI18 Wikipedia は 4×10^{-4} 未満で最も整合的である．

定性的ケーススタディ　非ゼロスラック上位制約の確認から，CEFR-J は具体名詞（*zoo*）と抽象動詞（*yearn*, *widen*）の配置差，CompLex は固有名詞を易しいとみなす注釈傾向，CWI18 Wikipedia は科学形容詞（*zoological*）を易しいとみなす傾向，EVKD は古風語彙の顕在化が主因だった．スラックが特定領域に集中するなら資源側の難易度割当を見直し，稀語彙由来ならコーパス拡充が有効である．

コーパス間比較　OANC でも同様の色分け分析結果を表 6 に示す（制約 10 万，seed 99）．CEFR-J はフィクションと電話会話，CompLex は旅行ガイドと技術文書，CWI18 Wikipedia はフィクションとジャーナル文体を相対的に強化する．