LLM サプライザルと語彙豊富度指標を用いた日本語エッセイの母語識別分析

TONG FENGYI[1] 久野雅樹[1]
[1]電気通信大学大学院情報理工学研究科
t2530092@edu.cc.uec.ac.jp hisano@uec.ac.jp

概要

本研究では，大規模言語モデルが算出するサプライザルと語彙豊富度指標（MTLD）を用いて，日本語エッセイの書き手の母語（日本語，中国語，韓国語）を識別する手法を提案する．Rinna-3.6B，ELYZA-Llama2-7B，LLM-jp-13B の 3 つの LLM を用いて分析を行った．その結果，サプライザルは書き手の母語によって異なる分布を示し，特に中国語話者のテキストにおいては特有の漢字使用や固有名詞がサプライザル値を高める要因であることが明らかになった．また，サプライザルと MTLD の相関分析から，学習者は語彙の難度が上がると文脈の不自然さが増す一方，母語話者はその影響を受けにくいという構造的な違いが示唆された．

1 はじめに

自然言語処理技術の発展に伴い，第二言語習得研究においても機械学習を用いた学習者データの分析が進んでいる．特に，学習者の書いたテキストからその母語を推測する母語識別は，学習者の誤用傾向や言語転移を理解する上で重要なタスクである．従来の手法では n-gram などの表層的な特徴量が用いられてきたが[1], 近年では大規模言語モデル（LLM）の予測確率を用いた分析が注目されている[2] ．

本研究では, LLM におけるサプライザル（surprisal）という特徴に着目する．サプライザルとは，ある単語が出現する確率の対数負値であり，モデルにとって予測しにくい単語ほど高い値を示す．本研究の目的は，このサプライザルと語彙の多様性を表す MTLD（Measure of Textual Lexical Diversity）を組み合わせることで，日中韓の日本語エッセイコーパスにおける母語識別を行い，その有効性と日中韓の言語的特徴を明らかにすることである．

2 関連研究

2.1 第二言語のライティングにおける LLM サプライザル

Hu と Cong は，LLM サプライザルを用いて中国語学習者のエッセイの熟達度（A2 から C1 レベル）を分類する試みを行った[3]．この研究では 260 編のエッセイからなるコーパスを使用し，繁体字中国語に特化したモデル（Taiwan-LLM）が最も高い効果量（$\eta^2 = 0.23$）を示し，隣接する全ての熟達度レベルを有意に識別できたことを報告している．これにより，サプライザルが学習者のレベル判定に有効な特徴量であることを示している．

2.2 流暢性評価

田村らは，機械翻訳の流暢性フィルタリングにおけるサプライザルの有効性を検証したが，既存の評価指標との間に明確な相関は見られなかったと報告している[4]．これは，機械翻訳の生成文における不自然さと，学習者の母語干渉による不自然さの質的な違いを示唆している可能性がある．

本研究は，これらの知見を拡張し，サプライザルを単なる熟達度や流暢性の評価だけでなく，母語間の言語的特徴の差異の分析に応用するものである．

3 分析 1：サプライザルと MTLD を用いた母語の分類

3.1 使用データセット

本研究では，日本語（JP），中国語（CN），韓国語（KR）を母語とする大学生による日本語作文データ（JCK コーパス[5]）を使用した．各母語 60 件ずつで 3 つのテーマ（「故郷について」「晩婚化」「趣味」）の作文を収録している．非母語話者（CN, KR）は JLPT で N1 合格レベル（上級）である．テキストの長さはそれぞれ 2000 文字程度でそろっていて．サプライザル計算へ長さの影響を排除できる．

3.2 実験手順

3.2.1 LLM サプライザルの算出

入力テキストをトークン化し，LLM を用いて直前の文脈を与えられた条件下での次のトークンの出現確率 $P(w_i|w_1 \ldots w_{i-1})$を計算した．個々のトークンのサプライザル $S(w_i)$は Hu と Cong に従い，以下の式で定義される[1].

$$S(w_i) = -\log P(w_i|w_1 \ldots w_{i-1})$$

文章全体のサプライザルは，構成する全トークンのサプライザル値の平均として算出した．モデルの選定にあたっては，予備実験においてパラメータ数の少ないモデル（1.3 億パラメータ等）では未知語の影響により値が不安定であったため，本分析では以下の 3 つの大規模言語モデルを採用した．

1. Rinna-3.6B: rinna/japanese-gpt-neox-3.6b (36 億パラメータ) [6]
2. ELYZA-Llama2-7B: elyza/ELYZA-japanese-Llama-2-7b-fast-instruct (70 億パラメータ) [7]
3. LLM-jp-13B: llm-jp/llm-jp-13b-v1.0 (130 億パラメータ) [8]

3.2.2 語彙豊富度（MTLD）の算出

MTLD は，テキスト内で TTR（Type-Token Ratio）が特定の閾値（通常 0.72）を下回るまでのトークン列の平均長を算出する指標である[9]．つまり，同一語を繰り返さずにどれだけ長く文章を記述できるかを数値化したものであり，本研究における書き手の語彙レベルの測定に適している．

3.2.3 分類モデルの構築

算出された 3 つのモデルによる LLM サプライザルと MTLD の計 4 つの特徴量とし，機械学習アルゴリズムである Random Forest Classifier（ランダムフォレスト）を用いて，書き手の母語クラス（JP，CN，KR）の 3 クラス分類を行った．

3.3 結果と考察

本節では，サプライザルの分布，サプライザルと語彙豊富度との相関，および母語分類の結果について述べ，各母語話者の言語的特徴を考察する．

3.3.1 LLM サプライザルの全体的傾向と国別比較

3 つの LLM を用いたサプライザルの計測結果において，すべてのモデルで共通した傾向が確認された．母語別の平均サプライザル（低いほどモデルにとって予測しやすい＝系列として自然）は，低い順に以下の通りとなった（図 1 参照）．

- JP: 最も低い値を示した．これはモデルの学習データが日本語主体であるため，当然の結果と言える．
- CN: 最も高い値を示した．
- KR: JP より高く，CN より低い中間的な値を示した．

箱ひげ図による分布を確認すると，CN のサプライザルは値の平均値が高い様子が見て取れる．

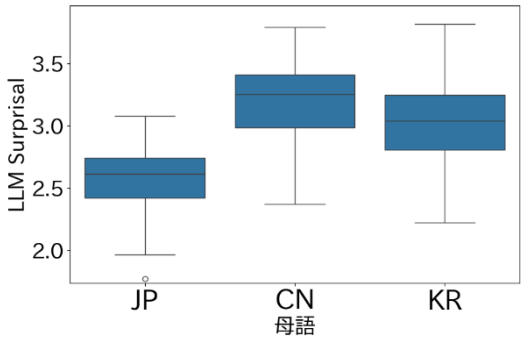

図 1　各国学習者のサプライザル分布図
—モデル LLM-jp-13B

CN の値が高くなる要因として，中国語母語話者は日本語に存在しない漢字や，日中で意味の異なる熟語（e.g.，「新聞」は中国語で「ニュース」を意味する）を使用する習慣があり，また固有の地名や専門用語を多用することが，モデルの予測難度を高めたと考えられる．

3.3.2 サプライザルと MTLD の二次元空間分布

書き手の言語特徴を視覚的に捉えるため，語彙豊富度を X 軸，LLM サプライザルを Y 軸とした二次元空間上に，各個人のスコアをプロットした．

空間上の分布傾向を見ると 3 つのいずれのモデルを用いた場合も，書き手の母語（JP, CN, KR）ごとに異なる分布領域（クラスター）が形成される様子が観察された．

- JP: 全体的に Y 軸サプライザルの低い領域に分布しており，MTLD の値にかかわらず予測難度が低い（自然である）ことが示された．
- CN: Y 軸の高い領域に広く分布しており，データのばらつきが大きい．これは前述の通り，漢字使用や固有名詞の影響によるものと考えられる．
- KR: JP と CN の中間的な領域に位置している．

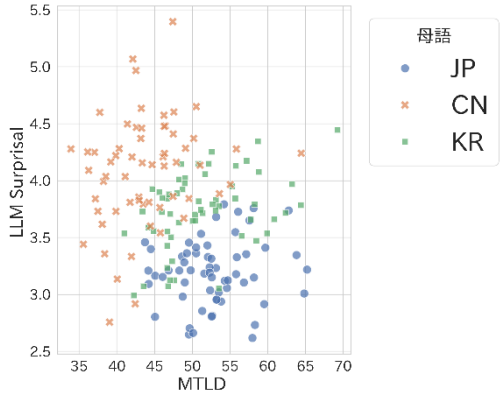

図 2　各母語話者のサプライザルと MTLD の散布図―モデル Rinna-3.6B

3.3.3 母語分類の結果

RFR モデルの分類正解率は 0.62 であった．これは，3 クラス分類におけるランダムな確率を大きく上回っており，3 つのモデルで算出されたサプライザルと MTLD というわずか 4 つの数値指標のみで，書き手の母語背景をある程度まで推定可能であることを示している．

さらに，混同行列の詳細な分析を行ったところ，クラス間の誤分類には非対称性が確認された．具体的には，JP は F 値が最も高く，他クラスとの分離が良好であった．一方で，KR と CN の間では一定の混同が見られた．特に，KR のデータが CN として誤分類されるケースよりも，CN のデータが誤分類されるケースの方が少ない傾向が見られた．これは，中国語話者のテキストが「漢字の多用」や「特有の固有名詞」という強い特徴（高いサプライザルと高い MTLD）を持っているため，他クラスと区別されやすいことを示唆している．逆に，サプライザルが中程度である KR は，特徴空間上で JP と CN の中間に位置するため，境界線付近での誤分類が生じやすかったと推察される．一方，JP はサプライザルが低く，特徴空間上で他クラスと明確に分離しているため，最も高い精度で識別された．

4 分析 2: 文単位のサプライザルおよび言語的特徴の分析

分析 1 で作文を単位とした分析によって観察された母語別の差違が具体的にどのような言語現象に起因するのかを明らかにするため，分析 2 では文単位での検討を行う．

4.1 使用データセット

分析 1 で使用した JCK コーパスの全文章を，句点などで区切り文単位に分割したデータを分析対象とした（表 1 参照）．

4.2 実験手順

分割されたすべての文に対して 3 つのモデルで LLM サプライザルを計算し，値が最も高い（予測困難な）上位の文を抽出した． さらに，各文に含まれる品詞（名詞，動詞，形容詞，副詞，助詞）の割合および漢字比率を算出し，母語ごとの平均値を比較した．

4.3 結果と考察

4.3.1 高サプライザル文の定性分析

サプライザル値が極めて高い文には以下の特徴が見られた（表 1 参照）．

- 固有名詞の多用 :「扎?（Zhalong）自然保護区」「綏化市」「杜尔伯特（Durbet）」など，中国の特定の地名や名称が含まれる文で著しく高い値（10.06, 9.14, 8.33）が記録された．これらは日本語の学習データに含まれる頻度が低いため，モデルが高いサプライザルを出力したと考えられる．
- 不自然なコロケーション :「満足やすい（8.36）」「もう出産できないものは類ものがあります（8.95）」のような，文法的な誤りや不自然な

語の結びつきもサプライザルを高める要因となった．

表 1　サプライザルが高い文の例（モデル Rinna-3.6B）

surprisal	sentence
10.05	「扎?」自然保護区の中心地にあった。
9.14	1982 年はじめて綏化市が設立した。
8.95	もう出産できないものは類ものがあります。
8.36	満足やすい。
8.33	次は杜尔伯特モンゴル自治県のモンゴル族の文化を感じられる。

4.3.2 言語特徴比率の母語別比較

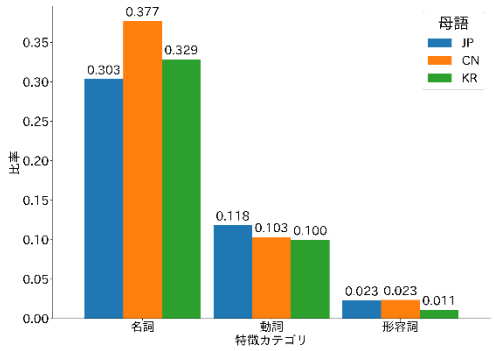

図 3　名詞・動詞・形容詞の比率

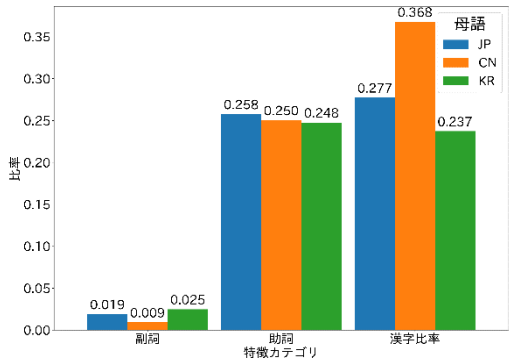

図 4　副詞・助詞・漢字の比率

各母語話者の作文における言語特徴の出現比率から以下の差異が明らかになった（図 3，4 参照）．

母語による差違が特に明確なのは名詞比率（図 3 参照）と漢字比率（図 4 参照）である．

名詞比率 CN の名詞比率 37.7%と最も高く，名詞（特に固有名詞や熟語）の多用が文章の予測難度を上げていることが示された．

漢字比率：中国語話者（CN）の漢字比率は 36.8%と突出して高く，日本語話者（JP）の 27.7%，韓国語話者（KR）の 23.7%を大きく上回った．これが CN のサプライザルを高める主要因の一つであると推察される．

文単位の分析により，サプライザルは単なる「作文の不自然さ」だけでなく，「母語に由来する固有名詞の使用」や「漢字使用頻度の高さ（特に CN）」に強く反応することが確認された．これにより，サプライザルは学習者の母語特性を捉える有効な指標となり得ることが示された．

5 結論と今後の課題

本研究では，LLM が算出するサプライザルと語彙豊富度指標 MTLD を組み合わせた手法の有効性を検証した．Rinna-3.6B，ELYZA-7B，LLM-jp-13B を用いた実験の結果，サプライザルは学習者の母語（特に中国語）に由来する特異な言語特徴を鋭敏に反映し，MTLD と組み合わせることで 62%の正解率で母語を識別可能であることを示した．特に，中国語話者のテキストにおける漢字の多用や固有名詞の影響が，サプライザルによって示される LLM の予測難度を著しく高める要因であることが明らかになった．

また，サプライザルと MTLD の相関分析からは，学習者は語彙の難易度が上がると文脈の不自然さが増す傾向が見られたのに対し，母語話者は語彙が豊富になっても自然さを保つことができるという，言語運用能力の構造的な違いも示唆された．

今後の課題としては，以下の点が挙げられる．

第一に，より大規模かつ高性能なモデル（GPT-4 や Llama-3 等）を用いた場合のサプライザル特性の検証である．モデルの性能向上により，学習者の微細な誤りに対する感度がどのように変化するかを明らかにする必要がある．

第二に，統語構造情報の活用である．本研究では文全体の平均サプライザルを用いたが，係り受け解析などを併用し，特定の品詞や構文構造における局所的なサプライザルを分析することで，より詳細な言語転移のメカニズムの解明に繋がると考えられる．

参考文献

[1] Scott Jarvis, Yves Bestgen and Steve Pepper. Maximizing Classification Accuracy in Native Language Identification. **Proceedings of the Eighth Workshop on Innovative Use of NLP for Building Educational Applications**, pp. 111-118, 2013.

[2] Wei Zhang and Alexandre Salle. Native Language Identification with Large Language Models. arXiv preprint arXiv:2312.07819, 2023.

[3] Jingying Hu and Yan Cong. Modeling Chinese L2 Writing Development: The LLM-Surprisal Perspective. **Proceedings of the Workshop on Cognitive Modeling and Computational Linguistics**, pp. 172–183, 2025.

[4] 田村 鴻希，土井 惟成，西田 直人，Junjie Chen，谷中 瞳．サプライザルを利用した日本語の流暢性フィルタリングの試み．言語処理学会 第29回年次大会 発表論文集，2023.

[5] JCK 作文コーパス，http://nihongosakubun.sakura.ne.jp/corpus，（2025-09 閲覧）．

[6] rinna Co., Ltd. rinna/japanese-gpt-neox-3.6b-instruction-ppo, https://huggingface.co/rinna/japanese-gpt-neox-3.6b-instruction-ppo,（2025-11 閲覧）．

[7] ELYZA, ELYZA-japanese-Llama-2-7b-fast-instruct, https://huggingface.co/elyza/ELYZA-japanese-Llama-2-7b-fast-instruct,（2025-11 閲覧）．

[8] llm-jp, llm-jp-13b-v1.0, https://huggingface.co/llm-jp/llm-jp-13b-v1.0,（2025-11 閲覧）．

[9] Philip M. McCarthy and Scott Jarvis. MTLD, vocd-D, and HD-D: A Validation Study of Sophisticated Approaches to Lexical Diversity Assessment. **Behavior Research Methods**, Vol. 42, No. 2, pp. 381–392, 2010.