N-011

語種によって単語の難易度が変わるか
Does Word Difficulty Vary With Word Type?

鄭 弯弯†
Wanwan Zheng

1. はじめに

日本語は世界の言語の中でも特に複雑で，習得が難しいといとされている．その一因として，日本語は複数の表記体系（漢字，ひらがな，カタカナ）を共時的に混用していることが挙げられる．異なる機能を持つ文字体系の混用は，異なる語種と結びつく．例えば，日本語固有の語は和語であり，漢語からきたものは漢字，欧米からきた借用語はカタカナで表記される．さらに，二つ以上の語種の組み合わせは混種語となる．使用する文字の種類によって印象が異なるため，同じ意味の伝達であっても，語種の扱いは使用者や目的によって変わることがある（菊池，1990; 垣花，2023）．特に近年，外来語の使用が若い世代に増加しており，世代間のコミュニケーションにおいて障害となる社会的な問題となっている（相澤，2007）．

一方で，単語の難易度は，単語の出現頻度，親密度，長さなどさまざまな側面との関係は検討されているが，日本語の特色である語種が単語の難易度に与える影響は量的に示されていない．そこで，本研究では，日本語を対象にして，膨大なコーパスを用いて単語の難易度と語種の関係を検討し，語種によって単語の理解に差が生じるかを明らかにする．さらに，語種の使い方は内容に影響される可能性があるため，内容の影響も考慮する．

2. 使用データ

データとして，「日本語教育語彙表」（Sunakawa ら，2012）を使用した．この語彙表は，「現代日本語書き言葉均衡コーパス（BCCWJ）」や日本語教科書のコーパスの語彙調査を行うことによって選定された 17,920 項目の見出し語からなっている．各見出し語には，経験豊富な日本語教師の主観判定に基づく 6 段階の難易度（初級前半・後半，中級前半・後半，高級前半・後半）の他，旧日本語能力試験の等級，品詞・語種・表記・アクセントなどの情報や，語義と用例（作例とコーパスからの実例）が搭載されている．難易度の分布は図 1 に示す．語種には，和語，漢語，外来語，混種語以外に定型句もあるが，図 2 に示されているように 25 個しかなく，他の語種の単語数よりはるかに少ないため，分析対象から除外した．

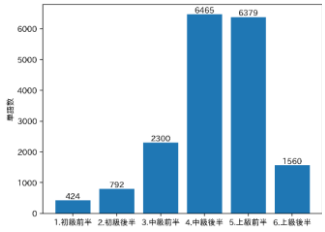

図 1.単語の難易度の分布

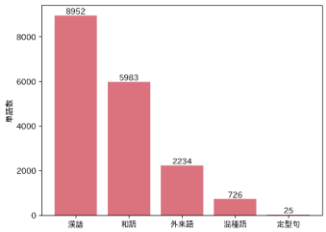

図 2.単語の語種の分布

†名古屋大学大学院人文学研究科　Graduate School of Humanities, Nagoya University

3. 分析

3.1 単語の難易度と語種の対応関係

まず，単語の難易度ごとの語種の分布を求めた（表 1）．難易度によって，和語と漢語の割合に変化があったが，すべての難易度において，和語と漢語の割合が最も高かった．詳細について，中級以上では和語より漢語のほうが多かったが，上級後半では和語と漢語の頻度はほぼ同程度であった．この結果から，漢語が和語より難易度が高いことを示唆された．外来語の割合は次に高く，その後は混種語が続いた．これは，BCCWJ 語種構成表（短単位）に示される和語，漢語，外来語，混種語の順序とほぼ一致している．

表 1. 難易度ごとの語種の割合

	和語	漢語	外来語	混種語
1.初級前半	**0.5307**	0.3325	0.0755	0.0401
2.初級後半	**0.4508**	0.2790	0.2096	0.0530
3.中級前半	0.3543	**0.4148**	0.1891	0.0404
4.中級後半	0.2877	**0.5460**	0.1341	0.0319
5.上級前半	0.3245	**0.5415**	0.0919	0.0415
6.上級後半	**0.4205**	0.4179	0.0949	**0.0660**

続きまして，難易度における語種の分布が異なるかどうかを検証するため，表 1 に対して，カイ二乗検定を行った．その結果，カイ二乗値は 0.44，p 値は 1.00 であり，単語の難易度と語種の分布は独立していると結論付けられた．難易度ごとの語種の分布に有意な差があると言えなかった．これに踏まえ，難易度ごとの語種分布に対してコサイン類似度を求め，具体的な類似度を図 3 に示す．すべての類似度は 0.86 以上であり，特定の語種は単語の難易度を支配することが見られなかった．

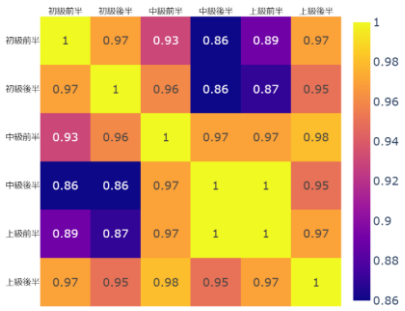

図 3.語種の分布による単語の難易度のコサイン類似度

3.2 単語の語種によって難易度の変化

同じ意味を持つが，語種が異なる単語ペアを抽出するため，言語モデル hottoSNS-w2v （松野ら，2019）を使用して各単語の分散表現を求めた．hottoSNS-w2v は日本語超大規模 SNS＋Web コーパスによる単語分散表現モデルであり

（登録語彙数は約 200 万語である），適切な前処理と長単位志向の分かち書きを行ってあるため，語彙情報が豊富かつ応用向きであると報告されている．

一方，「日本語教育語彙表」に含まれるが，hottoSNS-w2v に登録されていない語はベクトル化できないため，共通語を抽出し，更に重複語を削除した後，最終的に 16,866 語（16,866/17,920=94.12%）を抽出できた．

16,866 語から同じ意味を持つ語種が異なる語をペアで抽出するために，以下の処理を行った．

1. 抽出した各単語w_iに対して，他のすべての単語とのコサイン類似度を求め，最も類似している単語w_i'を判定した．
2. 類似単語ペア(w_i, w_i')の中から，順序を無視し，また語種が異なる条件で 2,684 ペアを抽出した．
3. 意味が同じペアの抽出を日本語 WordNet 同義体データベースを使用して行った．最終的に 409 ペアが得られた．

抽出した単語ペアの頻度と例を表 2 に示す．全体的に，漢語と和語のペアが最も多かった．これは，他の語種より，漢語と和語の数は元々数多いことが影響していると考えられる．続いて，漢語と外来語（130 ペア），和語と外来語（37 ペア），漢語と混種語（36 ペア）があった．外来語は漢語の意味との重複が明らかに多かった．

表 2. 同じ意味情報を持つ語種が異なる単語ペア

	頻度	例
(漢語，和語)	177	(多種多様，様々), (尺度，物差し)
(漢語，外来語)	130	(知能指数，IQ), (じゅうたん，カーペット)
(和語，外来語)	37	(物語，ストーリー), (まめ，キャンディー)
(漢語，混種語)	36	(浪費，無駄使い), (書店，本屋)
(和語，混種語)	24	(わがまま，自分勝手), (かわいそう，気の毒)
(混種語，外来語)	5	(台所，キッチン), (誕生日，バースデー)

図 4 は，各単語を語種と難易度に変換した後に描かれた共起ネットワーク図である．エッジの太さは，Simpson 係数を示しており，Simpson が高いほどエッジが太くなる．その結果，ほぼ同程度難易度の語種のペアが多かったが，例外もあった．上級前半の漢語と上級後半の外来語，初級前半の混種語は中級前半の外来語，中級後半の漢語と上級前半の和語のペアも目立った．

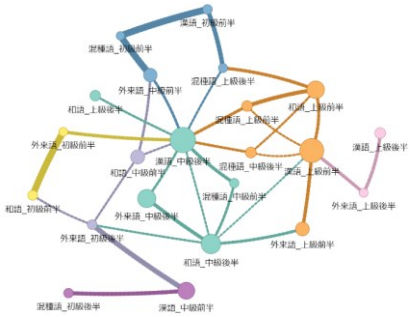

図 4.同じ意味を持つ単語の共起ネットワーク
(Simpson 係数>0.27;0.27 は Simpson 係数の平均値)

3.3 内容の影響を考慮する

ここまでの分析は，単語の意味を無視し，属性（単語の難易度，語種）のみを考慮していた．その結果，単語の語種によって難易度に差が生じないことが示された．一方で，樺島（1981）は，文章独自の内容を形作る要素となるものは，文章の内容に直接な関係を持ち，意味がはっきり限定された語であることを述べている．Halliday and Matthiessen (2004)は，機能文法の視点から，文章の目的やコンテキストに応じて語種がどのように変わるかを説明している．語種と意味分野との関係を調査する研究は以前から存在しているが，研究方法は，意味分野を「抽象的関係」，「人間活動」，「自然現象」などのように定義し，割合を統計するアプローチが一般的である．本研究では，単語の意味情報を捉える分散表現を使用し，意味情報を加味して単語の難易度と語種の関係を検討する．

単語の意味情報をまとめるため，16,866 語の分散表現を用いて，まず独立成分分析で五つの独立成分を抽出した．ソートした各独立成分に属する単語のスコア分布を図 5 に示す．独立成分スコアの絶対値が大きい単語は，その成分の特徴語になるため，閾値（独立成分 0~3： >0.02；独立成分 4：>0.0128）を設けて両端の単語を抽出した．その結果，独立成分 0～4 に対して，それぞれ 916，728，926，957，782，計 4,309 個の特徴語を抽出した．図 6 は t-SNE による特徴語の可視化であり，各独立成分がそれぞれの意味を持っていることが読み取れる．

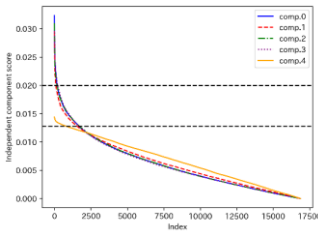

図 5. ソートした各独立成分に属する単語のスコア分布

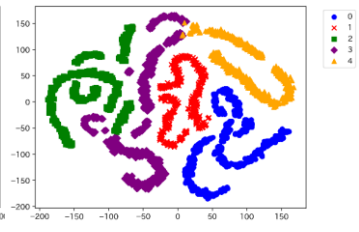

図 6. t-SNE による独立成分の可視化

続きまして，各独立成分の 4 つの語種におけるそれぞれの 5 つの難易度の割合を統計し，主成分分析を行った（図 7）．独立成分と関係なく，外来語・和語，混種語と漢語三つのグループが形成された．また，混種語漢語は中級後半と上級前半に位置し，混種語の難易度が最も高く，和語と外来語の難易度が最も低いことがわかった．

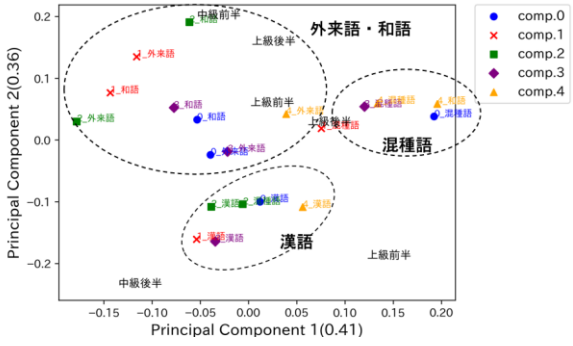

図 7. 各独立成分の 4 つの語種の主成分分析

4. おわりに

今回の分析で，和語と外来語の使用は内容に影響されるが，単語が持つ意味情報と関係なく，混種語の難易度が最も高く，続いては漢語，外来語・和語であることがわかった．

参考文献

[1] 菊池 悟，“語種イメージの分析—大学生と小学生の調査より—”, 岩手大学教育学部付属教育工学センター, No.12, pp. 67-79 (1990).