技術資料

Bayesian Linear Mixed Model による 単語親密度推定と位相情報付与

浅原　正幸 †

本論文では『分類語彙表増補改訂版データベース』に対する単語親密度推定手法について述べる．分類語彙表に収録されている 96,557 項目に対する評定情報を Yahoo! クラウドソーシングを用いて収集した．1 項目あたり最低 16 人（異なり 3,392 人）の研究協力者に，内省に基づいて「知っている」「書く」「読む」「話す」「聞く」の評定情報付与を依頼した．研究協力者の評定情報から単語親密度をベイジアン線形混合モデルにより推定した．また，推定された単語親密度と分類語彙表の語義情報との関連性について調査した．

キーワード：単語親密度，ベイジアン線形混合モデル，クラウドソーシング，位相情報

Word Familiarity Rate and Register Type Estimation Using a Bayesian Linear Mixed Model

Masayuki Asahara†

This paper presents research on word familiarity rate estimation using the 'Word List by Semantic Principles'. We collected rating information on 96,557 words in the 'Word List by Semantic Principles' via Yahoo! crowdsourcing. We asked 3,392 subject participants to use their introspection to rate the familiarity and register information of words based on the five perspectives of 'KNOW', 'WRITE', 'READ', 'SPEAK', and 'LISTEN', and each word was rated by at least 16 subject participants. We used Bayesian linear mixed models to estimate the word familiarity rates. We also explored the ratings with the semantic labels used in the 'Word List by Semantic Principles'.

Key Words: *Word Familiarity Rate, Bayesian Linear Mixed Model, Crowd-sourcing, Register*

1　はじめに

辞書は言葉に関するさまざまな特徴を集積したものである．発音・形態論情報・品詞・単語分類・統語情報・意味情報・位相・語源・語釈などにより整理される．単語の使用実態に基づく言葉の特徴として**単語親密度**がある．単語親密度は，人々がどのくらいその単語を知っているのか・使うの

† 人間文化研究機構国立国語研究所, NINJAL, Japan

かといった，人の主観的な評価に基づく指標である．NTT コミュニーケーション科学基礎研究所による『日本語の語彙特性』(天野，近藤 1999) は，単語親密度を含む情報を『新明解国語辞典第四版』の見出し項目約 80,000 語について付与した．また同データは朝日新聞の 1985 年から 1998 年の 14 年分の記事データにおける頻度情報も含む．しかしながら，評定情報の収集や頻度情報が 20 年以上前のものである．

本研究では，最近の単語親密度を評定することを試みる．日本語のシソーラスである『分類語彙表増補改訂版』(国立国語研究所 2004) の電子化データ『分類語彙表増補改訂版データベース』（以下「分類語彙表 DB」と呼ぶ）の語彙項目 94,838 語を対象に，単語親密度付与を行った．評定値の収集にあたっては，「知っている」の観点のほか，生産過程 ⇔ 受容過程や書記言語 ⇔ 音声言語の位相情報を含めるために，「書く」「読む」「話す」「聞く」の 4 つの位相情報についても質問事項に含めた．安価に，そして，継続的に調査を行うためにクラウドソーシングにより評定値の収集を行った．しかしながら，「日本語の語彙特性」の調査のように研究協力者に対する統制などに制約があり，研究協力者の個体差の影響を受ける問題がある．この問題を緩和するために，収集されたデータをベイジアン線形混合モデル (Bayesian Linear Mixed Model: BLMM) (Sorensen, Hohenstein, and Vasishth 2016) によりモデル化を行う．またシソーラスに単語親密度を付与することにより，統語分類・意味分類に基づく親密度・位相情報の評価もできるようになった．

本研究の貢献は以下の通りである：

- 日本語の大規模シソーラスに対する単語親密度情報の網羅的収集を行った．
- 単語親密度の評定にクラウドソーシングを用いた．
- 単語親密度の観点において「知っている」だけでなく，「書く」「読む」「話す」「聞く」の 4 つの位相情報についても検討し，単語の位相情報も評価した．これにより，生産過程 ⇔ 受容過程や書記言語 ⇔ 音声言語の対照比較ができる．
- 単語親密度の統計処理にベイジアン線形混合モデルを導入し，研究協力者の個体差の影響の軽減を行った．
- 語彙項目は分類語彙表 DB の見出し語を用いた．分類語彙表の統語・意味分類に対して親密度が推定できるほか，UniDic と分類語彙表の対応表 (近藤，田中 2020) と形態素解析器を用いて親密度を自動付与できる．さらに，『岩波国語辞典第五版』の語釈文と分類語彙表の対応表 (呉，近藤，森山，荻原，加藤，浅原 2019) の整備も進んでおり，語釈文との対照できる．

本稿の構成は以下の通りである．2 節では関連研究について示す．3 節ではクラウドソーシングに基づく単語親密度推定手法について示す．4 節で結果を示し，5 節にまとめと今後の展開について示す．

2 関連研究

『日本語の語彙特性』 (天野，近藤 1999) の単語親密度は，主観的な単語のなじみの程度とし，18 歳以上 30 歳未満の 40 人に対してアンケート調査により収集したものである．刺激を，「音声のみ」「文字のみ」「音声と文字の両方」の 3 つの手法で呈示し，1–7 の 7 段階評定を「できるだけまんべんなく」付与するように教示したうえで収集した．対象は『新明解国語辞典』第四版の見出し語 69,084 語とし，表記ゆれに基づき 88,569 パターンの刺激による．評定データは 1995 年 9 月から 1996 年 7 月にかけて，NTT 研究所内で収集された．機関内で収集したために，教示により統制してデータを収集しているが，公開評定データは 40 人の評定値の平均値による．研究協力者の個体差の影響を軽減するために，より洗練された統計処理が必要である．

また，『日本語の語彙特性』は，客観的な単語のなじみの程度として，テキストコーパスの出現頻度に基づく評定情報も含まれている．14 年分（1985–1998 年）の朝日新聞の新聞記事を形態素解析器『すもも』で解析したうえで，頻度情報をデータベース化した．Tanaka-Ishii and Terada (2011) は，単語親密度と大規模コーパス頻度情報との相関について調査した．水谷，河原，黒橋 (2018) は単語基本語情報を推定するために，各種語彙表のほか，ウェブコーパスの頻度情報を用いた．

岡久，久保，水谷，河原，黒橋 (2019) は，単語の定義（語釈）－被定義（見出し語）関係を用いて，単語の基本度の推定を行った．語釈文の収集にクラウドソーシングを用いた．11,936 語（被定義語）を対象として 1 語あたり 10 人分の語釈文を収集し，基礎データとした．定義－被定義関係は，シソーラスへの写像 (正津，白井，徳永，田中 2001) に用いられる．また，単語の基本度を数値化する基本的な考え方は野呂，徳田 (2007) による．

水谷，河原，黒橋 (2019) は，『JUMAN++』 の単語辞書に掲載されている単語 26,000 語を対象にクラウドソーシングを用いて，「習得時期」を「小学生になる前」「小学校低学年」「小学校高学年」「中学生になった頃」「単語の意味を知らない／見聞きしたことがない」の 5 段階で収集した．1 単語あたり 20 人の評定値を収集し，平均値により単語難易度データベースを構築した．

3 手法

本節では，単語親密度・位相情報の収集方法について示す．はじめに，収集対象の語彙項目の母集団である分類語彙表 DB について示す．次に，質問紙の構成について示す．最後に，統計処理手法について示す．なお，本手法は 2017 年に，形容詞・副詞・連体詞のみを対象として，全体の 10 分の 1 の規模で実行可能性を調査した (浅原 2017) うえで，再設計したものである．

3.1 分類語彙表の分類番号の例

『分類語彙表』は現代日本語を対象としたシソーラスで，岩波国語辞典の語彙項目から約 30,000 語が選ばれて採録された．1964 年に初版 (国立国語研究所 1964) が公刊された．その後，2004 年に増補改訂版が公刊され，増補改訂版の CSV ファイル（分類語彙表 DB）が研究用途に公開された [1].

分類語彙表 DB は，区切り文字を入れて 101,070 の語彙項目からなり，4 つの統語分類である類（体・用・相・他）と，階層的な意味分類が付与されている．各項目の分類は 5 ケタの数字による分類番号で示され，統語分類は 1 ケタ目，意味分類は小数点以下 4 ケタで表現する．意味分類の 1 ケタ目が部門，2 ケタ目までは中項目，4 ケタすべてが分類項目を表す．表 1 に，分類番号「1.1642」が割り当てられた「昨年」の例を示す．ここで，1 ケタ目「1」は，体の類を表す．小数点以下 4 ケタの「.1642」は階層的な意味分類を表し，1 ケタ目「.1」が部門「関係」を，2 ケタ目まで「.16」が中項目「時間」を，4 ケタ「.1642」が分類項目「過去」を表す．

3.2 質問紙の構成

本節では，単語親密度を推定するにあたって利用した，質問紙の構成について示す．質問紙は，Yahoo! クラウドソーシング上のフォームで構成する．クラウドソーシングを用いることにより，96,557 語 [2]に対して，短期間に低コストで評定情報を収集が可能である．まず，質問紙の構成として，書記刺激で呈示された語を知っているかどうか (KNOW) について確認する．音声刺激で呈示しない代わりに書記言語（書く (WRITE)・読む (READ)）・音声言語（話す (SPEAK)・聞く (LISTEN)）で利用されるかの観点を導入し，位相情報を推定する．さらに，生産過程（書く (WRITE)・話す (SPEAK)）・受容過程（読む (READ)・聞く (LISTEN)）の観点を導入した．以下に質問項目の 5 つの観点について示す：

KNOW: 知っている 単語の意味を知っていますか？

全く知らない (1)–(5) よく知っている

表 1 分類語彙表 DB

「昨年」：1.1642			
統語分類	意味分類		
類	部門	中項目	分類項目
体	関係	時間	過去
1.	.1	.16	.1642

[1] 商用利用は 200,000 円（税抜）.

[2] 分類語彙表 DB 101,070 項目のうち区切り文字 240 項目以外の 100,830 項目に対して調査を行ったが，クラウドソーシングの作業過程において 4,253 項目についてデータが得られなかった．得られなかった 4,253 項目については 2019 年 11 月に調査を行った．

WRITE: 書く　どのくらい普段書いているものに出現しますか？

全く出現しない (1)–(5) よく出現する

READ: 読む　どのくらい普段読んでいるものに出現しますか？

全く出現しない (1)–(5) よく出現する

SPEAK: 話す　どのくらい普段話すときに出現しますか？

全く出現しない (1)–(5) よく出現する

LISTEN: 聞く　どのくらい普段聞くときに出現しますか？

全く出現しない (1)–(5) よく出現する

図 1 に調査に用いた質問紙を示す．評定情報は，(1) 全く知らない／全く出現しない，(2) あまり知らない／あまり出現しない，(3) どちらともいえない，(4) 何となく知っている／たまに出現する，(5) よく知っている／よく出現するの 5 段階で収集するが，リッカート尺度としては収集しなかった．これは，2017 年の事前調査で数値だけ見て，全く反対の評定値を付与した研究協力者が散見されたため，それを回避するために行った．また，調査環境の制約により，天野，近藤 (1999)

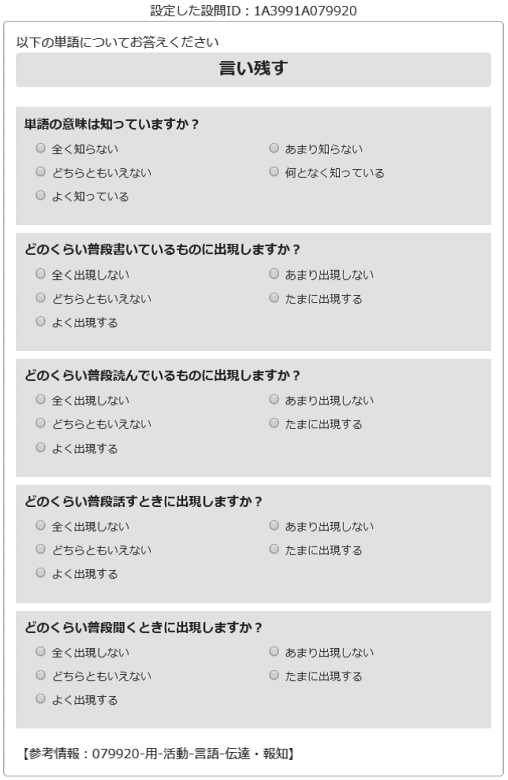

図 1　クラウドソーシング上のフォームで構成した質問紙の例

が実験室で行ったように音声刺激を呈示することは行わない．なお，本質問紙調査では，多義語の異なる語義についての評定（単語心象性）は行わない．研究協力者には「参考情報」として，質問紙の末尾に分類語彙表の分類項目を示す程度にとどめた．

本調査は異なりで 3,392 人の 20 歳以上の Yahoo! クラウドソーシングのアカウントを持っている方を対象に，2018 年 11 月に実施した．1 つの見出し語あたり少なくとも 16 回答を得た結果，有効データポイント数は 1,617,184 であった．収集の要した費用は 1,455,494 円であった．

3.3　統計モデル

収集した評定データは研究協力者ごと個体差がある．この個体差は，研究協力者の語彙力や回答の傾向に依存するバイアスである．この個体差をベイジアン線形混合モデルのランダム効果によりモデル化し，個体差の影響を軽減する．また，単語親密度情報も単語の特性としてランダム効果によりモデル化する．グラフィカルモデルを図 2 に示す．N_{word} は語彙項目 × 5 観点 (KNOW, WRITE, READ, SPEAK, LISTEN) の定義域，N_{subj} は研究協力者の数で，それぞれ単語・観点のインデックス $i : 1 \ldots N_{word}$，研究協力者のインデックス $j : 1 \ldots N_{subj}$ とする．$y^{(i)(j)}$ は語彙項目 × 5 観点 (KNOW, WRITE, READ, SPEAK, LISTEN) の値域で，y は平均 $\mu^{(i)(j)}$ 標準偏差 σ によって定義される正規分布とする：

$$y^{(i)(j)} \sim Normal(\mu^{(i)(j)}, \sigma).$$

σ は標準偏差としてのハイパーパラメータで，$\mu^{(i)(j)}$ は，切片 α と語彙項目 × 5 観点のランダム効果 $\gamma^{(i)}_{word}$ と研究協力者のランダム効果 $\gamma^{(j)}_{subj}$ の線形式で定義する：

$$\mu^{(i)(j)} = \alpha + \gamma^{(i)}_{word} + \gamma^{(j)}_{subj}.$$

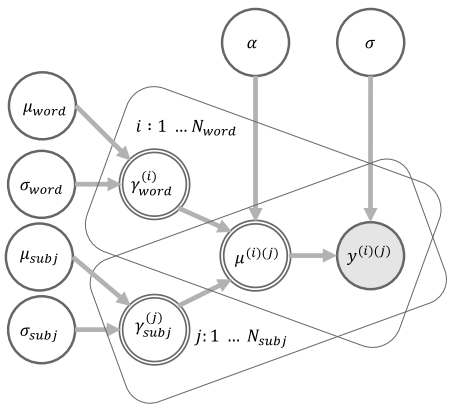

図 2　統計モデル

語彙項目 × 5 観点のランダム効果 $\gamma^{(i)}_{word}$ と 研究協力者のランダム効果 $\gamma^{(j)}_{subj}$ は，それぞれハイパーパラメータ平均 μ_{word}, μ_{subj}，分散 σ_{word}, σ_{subj} によって定義される正規分布によりモデル化する：

$$\gamma^{(i)}_{word} \sim Normal(\mu_{word}, \sigma_{word}),$$
$$\gamma^{(j)}_{subj} \sim Normal(\mu_{subj}, \sigma_{subj}).$$

このうち単語親密度はランダム効果 $\gamma^{(i)}_{word}$ の推定値である．一方，研究協力者の個体差はランダム効果 $\gamma^{(j)}_{subj}$ の推定値であるが，結果的に研究協力者の語彙力の評価値となる．ハイパーパラメータはデータから推定を試みた結果，収束しなかったために，平均 μ_{word}, μ_{subj} を 0.0，標準偏差 σ_{word}, σ_{subj} を 1.0 とした．

推定には R と Stan を用いた．warm-up 100 iteration のあと，5,000 iteration × 4 chains 並列でシミュレーションし，すべてのモデルは収束した．

4 推定した単語親密度・位相情報の分析

本節では，推定された単語親密度の質的評価を行う．4.1 節に，得られた 5 観点の評定値の分布と，研究協力者の個体差の分布について示す．4.2 節で各観点の上位 10 語・下位 10 語について確認する．4.3 節で分類語彙表の中項目（意味分類の第 2 レベル）による上位 10 カテゴリ・下位 10 カテゴリについて確認する．4.4 節にベイジアン線形混合モデルの効果について示し，4.5 節に『日本語の語彙特性』との比較を示す．4.6 節に結果のまとめを示す．

4.1 分布

図 3 に推定された単語親密度のヒストグラムを示す．x 軸が 単語親密度に相当する $\gamma^{(i)}_{word}$ で，y 軸がそのビン（ビンの幅 0.1）に入る頻度である．5 観点ごとに集計されており，KNOW が他の観点よりも高い単語親密度にある傾向がみられる．受容過程である READ と LISTEN が高く，生産過程である WRITE と SPEAK は低い傾向がみられた．言語の運用において，知識＞受容＞生産の順に語彙数が分布していることがわかる．また，値が大体 −2 から 2 に分布しており，おおよそ 5 段階評価の値としてそのまま利用できる．

図 4 に推定された研究協力者の個体差のヒストグラムを示す．x 軸が研究協力者の個体差 $\gamma^{(j)}_{subj}$ で，y 軸がビン（ビンの幅 0.1）に入る頻度である．標準正規分布でモデル化しているために，グラフもその形状になる．今回は利用しないが，この値は研究協力者の語彙力としてそのまま利用することができる．

他の分布に基づく個体差のモデル化については今後の検討課題とする．例えば，研究協力者の分布について標準正規分布とせずにモデル化を試みたが収束しなかった．今後，年次でデータを収集

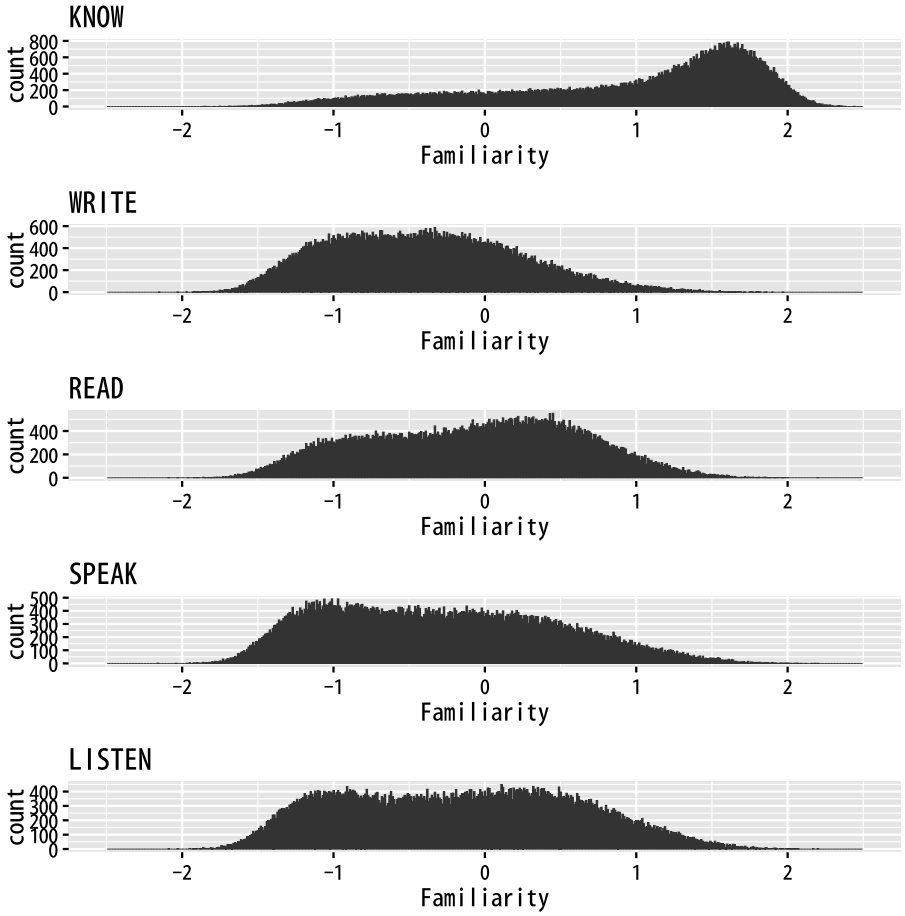

図 3 推定した単語親密度 ($\gamma^{(i)}_{word}$)：5 観点の分布

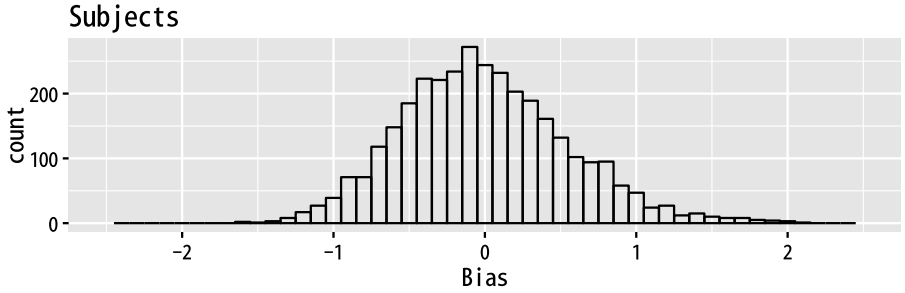

図 4 推定した研究協力者の個体差の分布 ($\gamma^{(j)}_{subj}$)

し拡充することで，他の分布に基づくモデル化を進めるとともに，調査年の要因を入れることで経年変化の調査を進めたい．

4.2 見出し語に対する評価

本節では，推定された単語親密度の上位 10 件・下位 10 件について確認する．

4.2.1 「知っている」

まず，最初に「知っている」(KNOW) の上位 10 件・下位 10 件について確認する．表 2 と 表 3 に上位 10 件・下位 10 件の事例を評定値と『国語研日本語ウェブコーパス』(NWJC)(Asahara, Maekawa, Imada, Kato, and Konishi 2014) の検索系『梵天』文字列検索 (浅原，河原，大場，前川 2018) のヒット件数とともに示す．検索時にはルビは削除した．上位 10 件には日常生活でよく利用する用語が出現し，下位 10 件には近年あまり使われない語が出現する．なお，活用語は基本形で調査したため，活用しない語よりも頻度が低い傾向にある．

梵天の文字列検索のヒット件数は活用語においては基本形のみの件数であり，低くなる傾向にある．さらにカタカナ語「スフ」は他の語の部分文字列になりやすいために件数が高くなる傾向がみられる．頻度情報との対照は，以上の注意が必要だが，基本的には高頻度語が上位に，低頻度語が下位に位置する．

また，WRITE, READ, SPEAK, LISTEN の単体の 4 観点についても基本的には同様の語彙がみられた．詳細な議論については，次小節以降の書記言語・音声言語もしくは生産過程・受容過程の観点による分析で示す．

表 2 「知っている」上位 10 件 (KNOW)

見出し語	KNOW	NWJC
全員	2.44	1,405,102
恋人	2.44	565,062
翌朝（よくあさ）	2.44	149,923
退社する	2.38	7,156
再会	2.38	635,900
本社	2.38	349,909
入社	2.37	380,872
人見知りする	2.36	6,338
持ち帰る	2.36	35,643
ストロー	2.36	166,713

表 3 「知っている」下位 10 件 (KNOW)

見出し語	KNOW	NWJC
うずみひ	−1.86	1
玉章（たまずさ）	−1.86	1,220
御稜威（みいつ）	−1.85	303
繞（にょう）	−1.85	2,325
鞅掌（おうしょう）する	−1.84	5
スフ	−1.82	769,698
驍名	−1.79	12
笈摺（おいずり）	−1.79	162
宇内（うだい）	−1.76	286
賢察	−1.75	1,094

4.2.2 書記言語・音声言語

次に，書記言語 (WRITE/READ) と音声言語 (SPEAK/LISTEN) の傾向を確認した．『日本語の語彙特性』の調査では，書記刺激と音声刺激の 2 つの刺激により統制するが，本実験では直接言語運用実態を質問することによる．ここでは 書記 – 音声 (WRITE + READ – SPEAK – LISTEN) の値を評価した．この値が正である場合に書記言語選好であり，この値が負である場合に音声言語選好である．

表 4 に書記言語選好上位 10 件（書記 – 音声 が正の値）を示す．記号関連の「アンパサンド」「句読点」や，手紙や文書で用いられる「上記」「追伸」「記」「前略」「下記」などがみられた．表 5 に音声言語選好上位 10 件（書記 – 音声 が負の値）を示す．「先っちょ」「バイバイ」「まんま」「どっこいしょ」など，話しことばが多くみられた．

4.2.3 生産過程・受容過程

本節では生産過程 (WRITE, SPEAK) と受容過程 (READ, LISTEN) の差について評価する．具体的には，生産 – 受容 (WRITE + SPEAK – READ – LISTEN) の正負により検討する．この観点は過去の研究では，管見の限り調査されていない．

表 6 に生産過程選好上位 10 件を示す．基本的には受容過程のほうが生産過程よりも親密度が高くなる傾向になるために，値は正であっても小さい．得られた語彙は，特定分野の専門用語が多かった．例えば，「毛管」「絆創膏」のような看護医療用語であったり，「吟詠する」など詩吟の用語であったりした．表 7 に受容過程選好上位 10 件を示す．「殺害」「書類送検」などの報道記事などで出現する語や，当時の大河ドラマの主人公「西郷隆盛」が上位にみられた．生産過程選好か受容過程選好かは，特定の分野で用いられる用語か，マスメディアで用いられる用語かを反映させていることがわかる．

表 4 書記言語選好上位 10 件

見出し語	書記 – 音声
上記	3.88
追伸	2.65
前述する	2.42
後述	2.35
記	2.30
前略	2.29
在中	2.18
アンパサンド［&］	2.17
句読点	2.12
下記	2.00

書記 – 音声: WRITE + READ – SPEAK – LISTEN

表 5 音声言語選好上位 10 件

見出し語	書記 – 音声
レジ袋	−3.07
先っちょ	−2.65
ちょろまかす	−2.59
バイバイ	−2.59
ヨーグルト	−2.52
ドライヤー	−2.47
まんま［その〜］	−2.46
それではまた	−2.42
鼻水	−2.42
どっこいしょ	−2.41

書記 – 音声: WRITE + READ – SPEAK – LISTEN

4.3 『分類語彙表』の分類に基づく評価

本節では分類語彙表の分類に基づく評価を行う．分類語彙表の中項目まで（小数点以下 2 ケタ）で推定した結果の平均を取り，値の上位カテゴリ・下位カテゴリについて分析を行う．

4.3.1 「知っている」

表 8, 表 9 に「知っている」分類語彙表の中項目カテゴリ上位・下位 10 件を示す．相・用の類が上位傾向にあり，体の類が下位傾向にある．「知っている」上位カテゴリは 3.53 (相-自然-生物) で「女性的」(KNOW=1.81),「男性的」(KNOW=1.71) などの単語が含まれる．「知っている」下位カテゴリは 3.52 (相-自然-天地) で「蕭条」(KNOW=-1.46) ,「巍巍」(KNOW=−1.35) などが含まれる．

表 6 生産過程選好上位 10 件

項目	生産 − 受容
毛管	0.76
物心（ぶっしん）	0.73
消却する	0.73
絆創膏	0.72
ふたとせ	0.71
揚げなべ	0.71
吟詠する	0.71
だるい	0.69
上辺（うわべ）	0.68
幽寂	0.66

生産 − 受容: WRITE + SPEAK − READ − LISTEN

表 7 受容過程選好上位 10 件

項目	生産 − 受容
送検する	−2.93
右翼	−2.71
書類送検	−2.69
巡業する	−2.59
西郷隆盛	−2.52
殺害（さつがい・せつがい）	−2.52
革命児	−2.48
護衛する	−2.47
識者	−2.42
再審	−2.41

生産 − 受容: WRITE + SPEAK − READ − LISTEN

表 8 「知っている」上位カテゴリ

中項目		KNOW
3.53	相-自然-生物	1.41
3.17	相-関係-空間	1.41
2.10	用-関係-真偽	1.35
3.56	相-自然-身体	1.34
2.56	用-自然-身体	1.32
2.14	用-関係-力	1.32
3.35	相-活動-交わり	1.32
4.32	他–呼び掛け	1.31
4.31	他–判断	1.29
3.57	相-自然-生命	1.26

表 9 「知っている」下位カテゴリ

中項目		KNOW
3.52	相-自然-天地	0.13
1.54	体-自然-植物	0.40
1.55	体-自然-動物	0.64
1.31	体-活動-言語	0.66
1.23	体-主体-人物	0.67
1.42	体-生産物-衣料	0.68
1.52	体-自然-天地	0.70
1.32	体-活動-芸術	0.71
4.50	他-動物の鳴き声	0.72
1.51	体-自然-物質	0.76

4.3.2 書記言語・音声言語

表 10, 表 11 に書記言語選好カテゴリと音声言語選好カテゴリを示す．体の類の活動・主体が書記言語選好である傾向がみられる．一方，他の類の呼びかけ・感動や相の類が音声言語選好である傾向がみられる．最も書記言語選好であるカテゴリは 1.31 (体-活動-言語) で，「上記」(WRITE + READ − SPEAK − LISTEN = 3.87)，「追伸」(WRITE + READ − SPEAK − LISTEN = 2.65) などの用語が見られた．もっとも音声言語選好であるカテゴリは 4.32 (他-呼びかけ) で「もしもし」(WRITE + READ − SPEAK − LISTEN = −1.75) が含まれる．

4.3.3 生産過程・受容過程

表 12, 表 13 に生産過程選好カテゴリと受容過程選好カテゴリを示す．

一般的に，受容過程の値 (READ, LISTEN) のほうが，生産過程の値 (WRITE, SPEAK) より

表 10 書記言語選好カテゴリ

中項目		書記 − 音声
1.31	体-活動-言語	0.13
1.32	体-活動-芸術	0.11
1.25	体-主体-公私	0.11
1.23	体-主体-人物	0.10
1.27	体-主体-機関	0.10
1.52	体-自然-天地	0.09
1.36	体-活動-待遇	0.08
2.31	用-活動-言語	0.07
1.53	体-自然-生物	0.07
3.52	相-自然-天地	0.07

書記 − 音声: WRITE + READ − SPEAK − LISTEN

表 11 音声言語選好カテゴリ

中項目		書記 − 音声
4.32	他-呼び掛け	−0.59
4.30	他-感動	−0.53
3.56	相-自然-身体	−0.44
2.56	用-自然-身体	−0.43
3.51	相-自然-物質	−0.42
3.18	相-関係-形	−0.33
3.50	相-自然-自然	−0.30
3.57	相-自然-生命	−0.29
4.50	他-動物の鳴き声	−0.29
1.43	体-生産物-食料	−0.28

書記 − 音声: WRITE + READ − SPEAK − LISTEN

表 12 生産過程選好カテゴリ

中項目		生産 − 受容
4.50	他-動物の鳴き声	−0.26
2.10	用-関係-真偽	−0.27
4.30	他-感動	−0.29
1.54	体-自然-植物	−0.30
4.32	他-呼び掛け	−0.30
3.52	相-自然-天地	−0.32
4.11	他-接続	−0.35
1.42	体-生産物-衣料	−0.35
1.55	体-自然-動物	−0.35
4.31	他-判断	−0.36

生産 − 受容: WRITE + SPEAK − READ − LISTEN

表 13 受容過程選好カテゴリ

中項目		生産 − 受容
1.27	体-主体-機関	−0.62
1.36	体-活動-待遇	−0.56
1.35	体-活動-交わり	−0.55
1.53	体-自然-生物	−0.54
3.17	相-関係-空間	−0.54
1.24	体-主体-成員	−0.54
2.35	用-活動-交わり	−0.53
2.36	用-活動-待遇	−0.53
2.34	用-活動-行為	−0.52
3.14	相-関係-力	−0.52

生産 − 受容: WRITE + SPEAK − READ − LISTEN

も大きい傾向にある．このため，生産 − 受容 (WRITE + SPEAK − READ − LISTEN) の値は，カテゴリレベルでは生産過程選好のものでも負になる．生産過程選好のカテゴリは 4.50 (他-動物の鳴き声) で 「げろげろ」 (WRITE + SPEAK − READ − LISTEN = 0.45) や「かーかー」(WRITE + SPEAK − READ − LISTEN = 0.23) が含まれる．受容過程選好のカテゴリは 1.27 (体-主体-機関) で「厚生労働省」(WRITE + SPEAK − READ − LISTEN = −2.23) や「金融庁」(WRITE + SPEAK − READ − LISTEN = −2.18) が含まれる．

4.4 ベイジアン線形混合モデルの効果

本節ではベイジアン線形混合モデルを用いた効果について概説する．

混合モデルを用いた効果として，得られた評価情報を稠密にすることがあげられる．図 5 に『日本語の語彙特性』の書記音声刺激の評定値の分布を，図 6 に統計処理前の評定値の分布（知っている）を，図 7 に統計処理後の評定値の分布（知っている）を，ビンの幅を 0.01 にしたヒストグラ

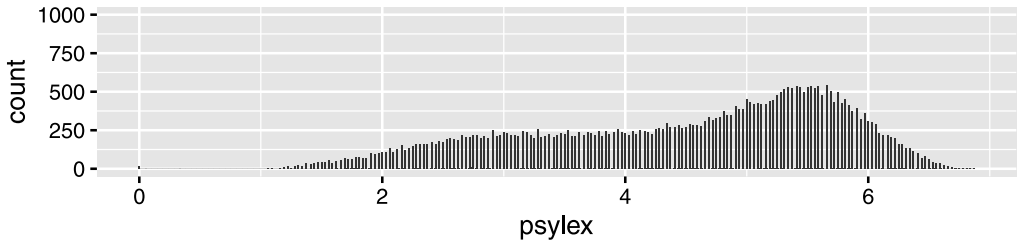

図 5 『日本語の語彙特性』の評定値の分布（書記音声刺激）

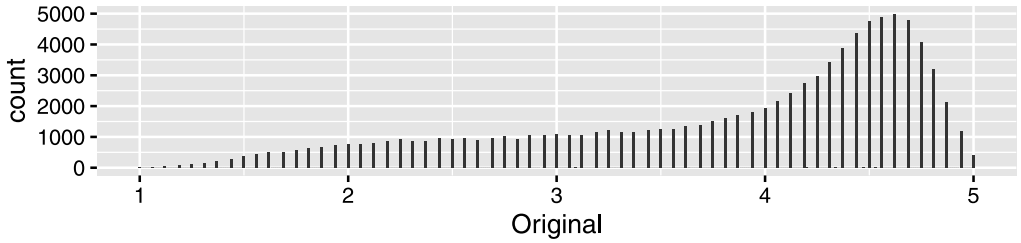

図 6 統計処理前の評定値の分布（知っている）

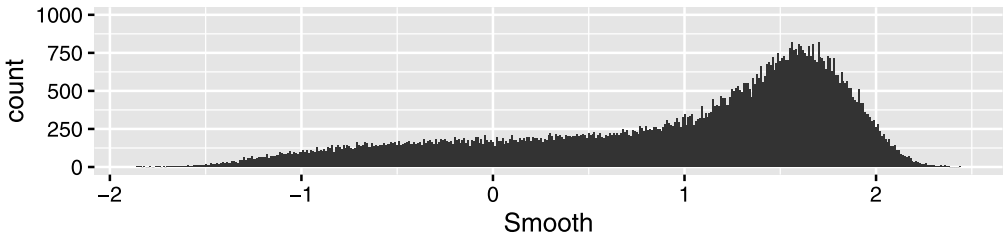

図 7 統計処理後の評定値の分布（知っている）

ムにより示す．『日本語の語彙特性』においては各語 40 人の評定値のため 1/40 = 0.025 単位の離散的な値となる．本データにおいても各語 16 人の評定値のため，統計処理をしなかった場合には 1/16 = 0.0625 単位の離散的な値となる．同じ評定値を持つ語が多く，事例がないビンも出現し，分布が疎になる（図 5，図 6）．例えば，評定値（知っている）が最高の 5.0 である項目は 410 事例あり，これらの同じ評定値が割り当てられている事例間の差異が得られない．一方，統計処理をした場合には，適切に平滑化が行われ，すべてのビンに項目が含まれ，分布が密になる（図 7）．

本研究ではランダム切片に基づく平滑化を行う．この平滑化においては，評定値の研究協力者ごとの平均値が用いられる．回答する評定値が高い研究協力者群に対しては，その研究協力者ごとの平均値に応じて評定値を減じ，回答する評定値が低い研究協力者群に対しては，その研究協力者ごとの平均値に応じて評定値を増じる．これにより，研究協力者の回答の個体差をモデル化したうえで，1 研究協力者内の語彙項目ごとの回答の差異に基づいて，語彙項目の評定値をモデル化することができる．他の平滑化手法として，ランダム傾きに基づく平滑化がある．この平滑化においては，研究協力者ごとの分散に応じて，評定値を増減する．

ベイジアン線形混合モデルは，この平滑化を含む線形モデルを，事後分布を生成するランダムサンプリングによりベイズ推定するものである．本研究ではランダム切片モデルのみならず，ランダム傾きモデルについても検討したが，収束しなかった．今後，データを増やして，よりあてはまりのよいモデルを検討したい．

4.5　『日本語の語彙特性』単語親密度との比較

本節では，『日本語の語彙特性』(天野，近藤 1999) の単語親密度との比較を行う．『日本語の語彙特性』と本データとで，表記と読みが同じ語彙項目が 41,634 対あった．『日本語の語彙特性』データにおいては，書記音声刺激・音声刺激・書記刺激の 3 つを比較対象とする．本データにおいては推定した 11 個の項目すべてを比較対象とする．本データは，表記と読みが同じであるが語義が異なる語彙項目がある．この場合，「知っている」の評定値が最も高いものを対照する．対照はスピアマンの順位相関係数に基づく．タイの場合は同順位の平均値を用いる補正を行う．表 14 に書記音声刺激に対する相関係数，表 15 に音声刺激に対する相関係数，表 16 に書記刺激に対する相関係数を示す．全ての検定において $p < 0.01$ であった．

本データの「書記-音声」「生産-受容」以外の指標は，『日本語の語彙特性』のいずれとも正の相関

表 14　『日本語の語彙特性』書記音声刺激と本データとの比較（順位相関係数）

知っている	書く	読む	話す	聞く	
0.861	0.837	0.857	0.863	0.870	
書記	音声	生産	受容	書記 – 音声	生産 – 受容
0.862	0.872	0.862	0.875	−0.435	−0.315

表 15 『日本語の語彙特性』音声刺激と本データとの比較（順位相関係数）

知っている	書く	読む	話す	聞く	
0.684	0.676	0.688	0.710	0.711	
書記	音声	生産	受容	書記 – 音声	生産 – 受容
0.694	0.715	0.704	0.710	−0.401	−0.218

表 16 『日本語の語彙特性』書記刺激と本データとの比較（順位相関係数）

知っている	書く	読む	話す	聞く	
0.850	0.831	0.850	0.855	0.861	
書記	音声	生産	受容	書記 – 音声	生産 – 受容
0.856	0.864	0.855	0.867	−0.424	−0.310

があることがわかる．そのなかでも書記音声刺激・書記刺激と強い正の相関があることがわかる．「書記-音声」「生産-受容」については中程度の負の相関がある．

4.6 結果のまとめ

まず，調査結果の分布をみると「知っている」＞「読む」「聞く」＞「書く」「話す」と，言語運用の負荷レベルで評定値に差異が出ることが明らかになった．

「知っている」「書く」「読む」「話す」「聞く」の 5 観点の親密度上位語・下位語を確認すると，ほぼ同じ語が分布することが確認できた．このため，書記言語 ⇔ 音声言語，生産過程 ⇔ 受容過程の 2 軸で分析することを試みた．具体的には対となる値を引き算することで，評定値の偏りがある語の上位 10 件について検討した．書記言語選好・音声言語選好の対照分析においては，書きことば・話しことばの語彙を適切にモデル化できていることを確認した．また，生産過程選好・受容過程選好の対照分析においては，特定分野の用語・マスメディアの用語といった比較ができることがわかった．さらに分類語彙表の中項目までのカテゴリ情報で同様の分析を試みた．用・相の類のほうが体の類よりも親密度が高い語が多いことがわかった．また体の類の活動・主体が書記言語選好で，他の類の呼びかけ・感動が音声言語選好であった．生産過程選好のカテゴリとして他の類が多くみられる一方，受容過程選好のカテゴリは体の類の主体・活動，用の類の活動など，マスメディアで用いられやすい語が多かった．

ランダム切片を用いたベイジアン線形混合モデルにより，同順位の語彙項目を多く含む疎な評定値を，研究協力者の個体差を考慮した密な評定値に変換できることを図 6，図 7 の対比により示した．また，先行研究の『日本語の語彙特性』との相関について検討した（表 14，表 15，表 16）．

5 おわりに

本稿では，分類語彙表 DB の見出し語に対する単語親密度付与について解説した．短期間で安価に作業を進めるためにクラウドソーシングを用いて，1 見出し語あたり 16 人の評定値を付与した．得られたデータをベイジアン線形混合モデルのランダム効果を用いて，親密度と研究協力者の個体差を同時にモデル化を行い，研究協力者の評定値のバイアスを吸収した．評定する観点として「知っている」だけでなく，「書く」「読む」「話す」「聞く」を含めた 5 観点を導入した．これにより，書記言語 ⇔ 音声言語・生産過程 ⇔ 需要過程の 2 軸の位相情報を導入する．軸の反対方向の観点との差分を取ることにより，言語使用に偏りがある語を抽出することができた．UniDic-分類語彙表対応表 (近藤，田中 2020) が整備されており，言語資源を組み合わせることで形態素解析結果に単語親密度を割り当てることも可能である．

構築したデータは `https://github.com/masayu-a/WLSP-familiarity/` で配布するほか，辞書検索ツール CradleExpress `https://cradle.ninjal.ac.jp/` で閲覧可能である．

以下，今後の課題について示す．本研究では，研究協力者の評定値の個体差を標準正規分布によりモデル化した．他のモデルについても検討したが，現状のデータポイント量では収束が困難であった．今後，年次でデータを拡充することで適切なモデル化をすすめたい．その際には調査年の要因を含めるとともに，語彙を UniDic の語彙素から表記ゆれを展開するなど表記間の差異についても検討したい．

また，本データの単語親密度，『現代日本語書き言葉均衡コーパス』に対する分類語彙表番号付与データ (加藤，浅原，山崎 2019)，岩波国語辞典の語釈文 (呉他 2019) の定義-被定義関係などを用いて，単語心象性の調査を行う．単語心象性は，多義語の語義ごとの親密度を推定するものである．刺激呈示時に語義の情報を結びつける必要がある．本研究では質問紙の末尾に参考情報として分類語彙表の分類項目の情報を示しているが，ほとんどの研究協力者が語義を意識していないと考える．用例や語釈文などの情報を用いて，明示的に語義を示すことで，語義ごとの親密度である単語心象性の調査を進めたい．

謝 辞

本研究は，国立国語研究所コーパス開発センター共同研究プロジェクト「コーパスアノテーションの拡張・統合・自動化に関する基礎研究」によるものです．本研究の一部は JSPS 科研費 基盤研究 (A) 17H00917，基盤研究 (C) 19K00591，基盤研究 (C) 19K00655，挑戦的研究 (萌芽) 18K18519，新学術領域研究 18H05521 の助成を受けたものです．

参考文献

天野成昭，近藤公久（編）(1999). 日本語の語彙特性. 三省堂, 東京.

浅原正幸 (2017). 『分類語彙表』に対する単語親密度推定―相の類を中心に―. 電気情報通信学会思考と言語研究会 (TL), **TL2017-22**, pp. 45–50.

浅原正幸，河原一哉，大場寧子，前川喜久雄 (2018). 『国語研日本語ウェブコーパス』とその検索系『梵天』. 情報処理学会論文誌, **59** (2), pp. 299–305.

Asahara, M., Maekawa, K., Imada, M., Kato, S., and Konishi, H. (2014). "Archiving and Analysing Techniques of the Ultra-Large-Scale Web-Based Corpus Project of NINJAL, Japan." *Alexandria*, **1–2**, pp. 129–148.

加藤祥，浅原正幸，山崎誠 (2019). 分類語彙表番号を付与した『現代日本語書き言葉均衡コーパス』の書籍・新聞・雑誌データ. 日本語の研究, **15** (2), pp. 134–141.

国立国語研究所 (1964). 分類語彙表. 秀英出版, 東京.

国立国語研究所 (2004). 分類語彙表増補改訂版. 大日本図書, 東京.

近藤明日子，田中牧郎 (2020). 「分類語彙表番号―UniDic 語彙素番号対応表」の構築. 国立国語研究所論集, **18**, pp. 77–91.

水谷勇介，河原大輔，黒橋禎夫 (2018). 日本語単語の難易度推定の試み. 言語処理学会第 25 回年次大会発表論文集, pp. 670–673.

水谷勇介，河原大輔，黒橋禎夫 (2019). クラウドソーシングを用いた習得時期の想起質問に基づく単語難易度データベースの構築. 言語処理学会第 25 回年次大会発表論文集, pp. 1503–1506.

野呂智哉，徳田雄洋 (2007). 語釈文記述のための日本語定義語彙の構築に関する一考察. 言語処理学会第 13 回年次大会発表論文集, pp. 626–629.

岡久太郎，久保圭，水谷勇介，河原大輔，黒橋禎夫 (2019). クラウドソーシングにより収集した語釈文を基にした単語の基本度推定. 言語処理学会第 25 回年次大会発表論文集, pp. 1499–1502.

正津康弘，白井清昭，徳永健伸，田中穂積 (2001). 国語辞典の語釈文の解析と語義のシソーラスへのマッピング. 第 15 回人工知能学会全国大会論文集, 2B2-01.

Sorensen, T., Hohenstein, S., and Vasishth, S. (2016). "Bayesian Linear Mixed Models Using Stan: A Tutorial for Psychologists, Linguists, and Cognitive Scientists." *Quantitative Methods for Psychology*, **12** (3), pp. 175–200.

Tanaka-Ishii, K. and Terada, H. (2011). "Word Familiarity and Frequency." *Studia Linguistica*, **65** (1), pp. 96–116.

呉佩珣，近藤森音，森山奈々美，荻原亜彩美，加藤祥，浅原正幸 (2019). 『分類語彙表』と『岩波国語辞典第五版タグ付きコーパス 2004』の対応表. 言語資源活用ワークショップ 2019 発表論文集, pp. 337–342.

略歴

浅原　正幸：2003 年奈良先端科学技術大学院大学情報科学研究博士後期課程修了．2004 年より同大学助教．2012 年より人間文化研究機構国立国語研究所コーパス開発センター特任准教授．2019 年より同教授．博士（工学）．言語処理学会，日本言語学会，日本語学会各会員．

（2019 年 7 月 31 日　受付）
（2019 年 9 月 19 日　再受付）
（2019 年 10 月 21 日　採録）