クラウドソーシングによる単語親密度データの構築 (2021 年版)

浅原 正幸
国立国語研究所
masayu-a@ninjal.ac.jp

概要

2018 年〜2021 年にかけて、継続して単語親密度調査を進めている。調査においては『分類語彙表』に出現する見出し語に対して「知っている」「書く」「読む」「話す」「聞く」の利用の程度をクラウドソーシングにより質問紙調査を行った。今回、４年間の累積データに対して、ベイジアン線形混合モデルを見出し語に基づき実施したので報告する。また同データベースの検索ツールについても紹介する。

1 はじめに

我々は 2018 年にクラウドソーシングによる単語親密度調査を開始した。最初の公開データ [1] にはデータ量が少なく、その後毎年継続調査を行いデータを増補してきた。また従前のモデル構築方法にも問題があった。本研究では増補したデータとともに新たに単語親密度データベースを再構築したので報告する。改変点では以下のとおりである：

- 欠損していたデータについても補完し、全見出しに単語親密度を付与した。
- 同一見出しについて、単一の単語親密度を付与するように変更した。
- 回答数が少ない実験協力者データを排除した。
- ベイズ主義的分析においては、単語親密度・実験協力者の語彙力をともに標準正規分布によりモデリングを行った。

また、分類語彙表を検索するツール CradleExpress 上に、単語親密度情報を格納した。分類語彙表番号に基づく語義階層ごとの単語親密度順ソートが可能になった。

以下ではデータの構築方法とともに検索ツールについて紹介する。

2 評定値の収集・モデリング

2.1 評定値の収集

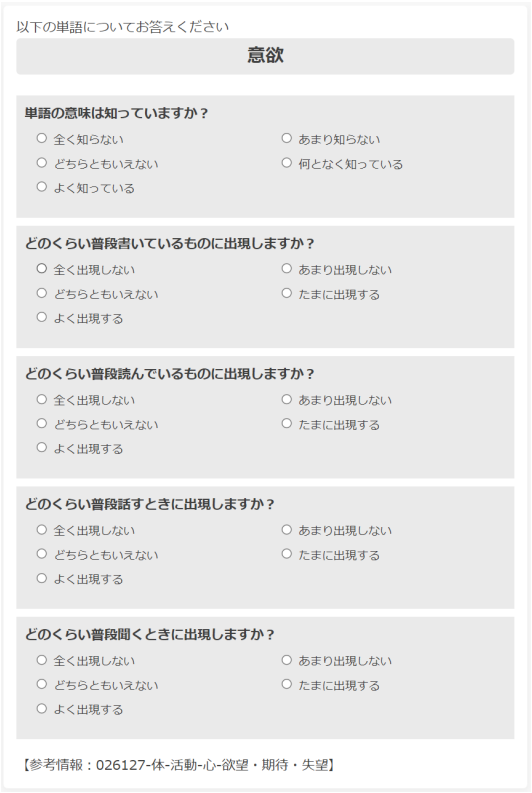

図 1 評定値収集画面

調査時期	実験協力者数	収集回答数
2018/11/15-21	3391	1617215
2019/11/14-22	2421	288000
2020/10/09-12	2372	943295
2021/09/12-14	2396	385380

表 1 収集評定値の統計

分類語彙表に登録されている「見出し」を刺激として、「知っている」(KNOW)・「書く」(WRITE)・「読む」(READ)・「話す」(SPEAK)・「聞く」(LISTEN) の 5 観点について Yahoo! クラウドソーシングを用いてアンケート調査を実施した。図 1 に調査に用いた評定値収集画面の例を示す。2018 年に初回の調

査を行ったのち、欠損値を補完するために 2019 年に未収集もしくは評定値の分散が高い見出しを中心に評定値の収集を行った。2020 年はクラウドソーシングによる読み時間データ収集 [2] において、実験協力者の語彙力を評定する事前調査として実施した。2021 年は区切り文字を除いた全見出しに対して調査を行った。表 1 に調査時期と実験協力者数・回答数を示す。

2.2 モデリング

見出し×5 観点	420570
実験協力者数（異なり）	6732
データポイント	15691265

表 2 分析対象データ

モデリングに際して、データの整理を行った。まず、実験協力者に呈示した見出しごと (84114 見出し) に KNOW, WRITE, READ, SPEAK, LISTEN の 5 観点に対して一意の ID (WID: 1〜420570) を付与し、そのランダム効果を単語親密度とする。従前のモデル [1] では、見出しが同じである異なる分類語彙表番号を持つ語（多義語）は、分類語彙表番号ごとに推定を行っていたが、今回は見出しごとに推定を行う。

収束性を担保するために 150 回答以上の実験協力者である異なり 6732 人分のデータのみを利用する。実験協力者に対して一意の ID (SID: 1〜6732) を付与し、このランダム効果も「実験協力者の語彙力」として用いる。また 1 実験協力者が 2 回以上回答した見出しについては排除した。結果 15691265 件のデータポイントを得た（表 2 ）。

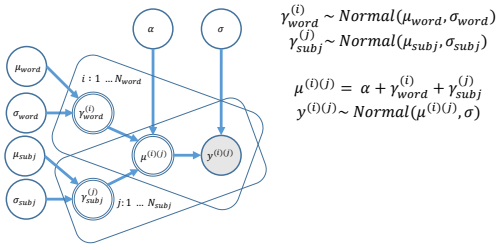

図 2 統計モデルの概要

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} は見出し×5 観点 WID の定義域、N_{subj} は実験協力者 SID の定義域である。それぞれ単語・観点のインデックス $i : 1 \ldots N_{word}$、実験協力者のインデックス $j : 1 \ldots N_{subj}$ とする。$y^{(i)(j)}$ は見出し×5 観点の評定値 (RATE) で、アンケート結果を 1（全く知らない／出現しない）〜5（よく知っている／出現する）に数値化したものを利用する。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 と見出し×5 観点のランダム効果 $\gamma^{(i)}_{word}$ と実験協力者のランダム効果 $\gamma^{(j)}_{subj}$ の線形式で定義する：

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

見出し×5 観点のランダム効果 $\gamma^{(i)}_{word}$ と実験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 $\mu_{word} := 0.0$, $\mu_{subj} := 0.0$、標準偏差 $\sigma_{word} := 1.0$, $\sigma_{subj} := 1.0$ によって定義される正規分布によりモデル化した。

$$\gamma^{(i)}_{word} \sim Normal(\mu_{word} := 0.0, \sigma_{word} := 1.0),$$

$$\gamma^{(j)}_{subj} \sim Normal(\mu_{subj} := 0.0, \sigma_{subj} := 0.5).$$

このうち単語親密度は見出し語×5 観点のランダム効果 $\gamma^{(i)}_{word}$ の推定値である。一方、実験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、結果的に実験協力者の語彙力の評価値となる。

推定には R と Stan を用いた。warm-up 30 iteration のあと、300 iteration × 3 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（$\hat{R} < 1$）。

3 データの分布

図 3 に得られた語彙力および単語親密度の分布を示す。語彙力は標準正規分布としてモデル化したためにこのような分布になっている。この語彙力のデータは匿名化したうえで読み時間評価実験 [2] に参加した方の属性値として利用する。

単語親密度の分布は、「知っている」>「読む」「聞く」>「書く」「話す」の順に大きい。「生産」は「書く」+「話す」、「受容」は「読む」+「聞く」、「書記」は「書く」+「読む」、「音声」は「話す」「聞く」の値である。「受容」が「生産」より大きい傾向があるほか、「書記」が「音声」より大きい傾向がみられる。

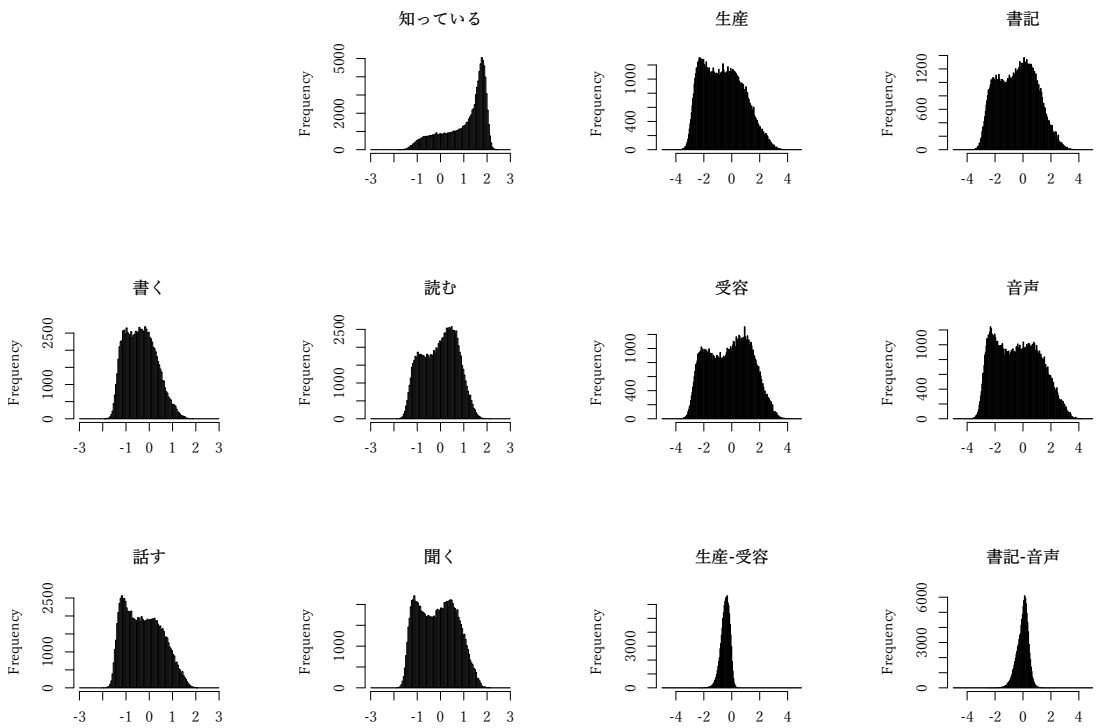

図 3　語彙力および単語親密度の分布

また、「生産-受容」は「書く」+「話す」-「読む」-「聞く」の値である。一般的に多くの語が「受容」>「生産」であるために負の値になるが、マスコミなどで利用される語で個人ではあまり利用しない語が「生産」>「受容」となり正の値を持つ場合がある。

最後に、「書記-音声」は「書く」-「話す」+「読む」-「聞く」の値である。正の値であれば書き言葉特有の語、負の値であれば話し言葉特有の語であることがわかる。

4　検索ツール: CradleExpress

次に分類語彙表の検索ツール CradleExpress [1] について紹介する。図 4 に検索画面を示す。

「詳細条件」で分類語彙表に基づく情報で絞り込み検索ができるほか、「親密度」でも絞り込み・並べ替えができる。以下では、分類番号 1.5501 「体-自然-動物-哺乳類」の親密度について確認する。図 5 に「知っている」上位 5 件の見出し語を示す。「知っている」列で並び替えを行うと、動物園などで観られる「ゴリラ」・「コアラ」・「やぎ」・「ライオン」が上位を占めた。また、分類学上の「哺乳類」が 3 位であった。

次に「生産」(「書く」「話す」を足したもの)・「受容」(「読む」・「聞く」)列で並び替えを行うと、図 6,7 のようになった。身近な動物である「猫」「犬」「野良猫」や、食肉としての「牛」「豚」が上位を占めた。また、分類学上の「動物」が 2 位（受容）・3 位（生産）であった。

基本語彙の観点からは、「理解語彙」(知っている)と「利用語彙」(書く・読む・話す・聞く)の 2 つの観点で大きな差があることがわかる。「生産」と「受容」に大きな差がある語としては、「盲導犬」「猛虎」など特定の媒体の頻出する表現が確認された。

また、書き言葉で用いられるか、話し言葉で用いられるかを判別するために、「書記-音声」(「書く」+「読む」-「話す」-「聞く」)の評定値により、降順(書記優位)・昇順（音声優位）したものを、図 8,9 に示す。前者が特定の小説・書籍・新聞などの書き言葉で頻出する語である一方、後者が対話などの話し言葉で頻出する語である。

このような評定値は認知言語学における基本レベルカテゴリとなりうる典型例を示すことができる。しかしながら、認知的に基本的なものとして、「理解語彙」(知っている)を指し示すか、「利用語彙」(書く・読む・話す・聞く)を指し示すかにより、語

1)　https://cradle.ninjal.ac.jp/wlsp/

図 4　CradleExpress 検索要求指定画面

彙の分布が著しく異なる。さらに、「書記」（書く・読む）か「音声」（話す・聞く）かにより評定値差の大きなものがあることから、プロトタイプ性の認定にあたっては、そのレジスタ（書き言葉か話し言葉か）も含めて検討すべきことが示唆された。

5　おわりに

本稿では、継続的に調査を行っている単語親密度データの 2021 年時点の累積データに基づいて、見出し語単位で統計処理を行ったデータについて示した。また、同データを検索するツールについても紹介した。

本調査では、単語単位での評定を行ったために、多義語においては、その語義ごとの利用実態を明らかにすることができない。例えば、上の例の「なまけもの」は字義通りの「哺乳類」の評定値か「怠惰」の評定値かがわからない。そこで、コーパスの出現ごとの印象評定調査を行った [3]。今後、コーパス上の用例と本データとを対照し、ヒトの語彙のとらえ方を明らかにしていきたい。

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の成果物です。また、科研費 17H00917, 18K18519, 19K00591, 19K00655 の支援を受けました。

Surface	知っている (f00)	書記 (f07)	音声 (f08)
ゴリラ	2.19434031554613	0.31748989420890816	0.645253936266903
コアラ	2.16988931425696	0.27520523348496	1.115538591867071
哺乳類	2.16628019642134	0.649779029443949	0.975175398158484
やぎ	2.13665095336729	0.06419158274259999	0.2032152535215712
ライオン	2.1223051693044	0.3590675971468735	0.680990317360433

図 5　分類番号 1.5501 「知っている」上位

Surface	知っている (f00)	生産 (f05)	受容 (f06)
猫	2.09645247543686	2.4814139287023798	3.07674663434263
犬	2.03476651063054	2.342248021812839	2.81346352328105
動物	1.91034925178924	2.309858375301816	2.9280813084437503
牛	1.70435360708	1.594586907518292	1.791518462452047
野良猫	2.07294104928732	1.504459343886654	2.3108553332465904

図 6　分類番号 1.5501 「生産」上位

Surface	知っている (f00)	生産 (f05)	受容 (f06)
猫	2.09645247543686	2.4814139287023798	3.07674663434263
動物	1.91034925178924	2.309858375301816	2.9280813084437503
犬	2.03476651063054	2.342248021812839	2.81346352328105
野良猫	2.07294104928732	1.504459343886654	2.3108553332465904
豚	1.92754361748217	1.297582058297013	2.0105191083476592

図 7　分類番号 1.5501 「受容」上位

Surface	知っている (f00)	書記-音声 (f10)
野うさぎ	1.64525871945305	1.031844155596917
やまあらし	1.17036381793061	0.7979713163135049
赤馬	0.308798738053768	0.7354525701993202
霊長類	1.2689718203082	0.666646635471029
土佐犬	1.58893703003315	0.6509951531038392

図 8　分類番号 1.5501 「書記-音声」上位（書記優位）

Surface	知っている (f00)	書記-音声 (f10)
くま	2.00338712343832	-1.650574420602923
小犬	2.03645330177898	-1.0887542842789781
なまけもの	1.78461415901402	-1.0635196871282888
象	1.96598447183801	-1.037933400148984
子猫	2.07086346878817	-1.007287355772866

図 9　分類番号 1.5501 「書記-音声」下位（音声優位）

参考文献

[1]浅原 正幸. Bayesian linear mixed model による 単語親密度推定と位相情報付与. **自然言語処理**, 27(1):133–150, 2020.

[2]浅原 正幸. クラウドソーシングによる大規模読み時間データ収集. In **言語処理学会第** *27* **回年次大会発表論文集**, pages 1156–1161, 2021.

[3]加藤 祥 and 浅原 正幸. 『現代日本語書き言葉均衡コーパス』に対する印象評定情報付与. In **言語処理学会第** *28* **回年次大会発表論文集**, 2022.

A　単語親密度の検索手法

以下では、CradleExpress による単語親密度の検索手法について示す。CradleExpress は UniDic と分類語彙表が格納されており、このうち分類語彙表が格納されている https://cradle.ninjal.ac.jp/wlsp/ を用いる。

図 10　CradleExpress 検索結果：分類番号 1.5501 デフォルト画面

図 4 に示す検索要求指定画面に対して、検索したい「分類」を指定する。例えば、1.5501 を指定すると、図 10 のような画面を得る。

単語親密度が表示されるが、各列の「▲」「▼」をクリックすることで「昇順」「降順」の並び替えができる。または、検索要求画面下部にある「並べ替え」の「キー：」（図 11）により、調べたい単語親密度情報を指定する。

図 10 の表示結果は、右側にも「生産」「受容」などの評定値があり、横スクロールで右に移動して表示できる。または、図 12 のように「設定」をクリックし、カラム幅を調整することで、必要な列のみを表示できる。表示しない列は「0」に設定し、表示したい列は「auto」もしくは適切な値を設定する。

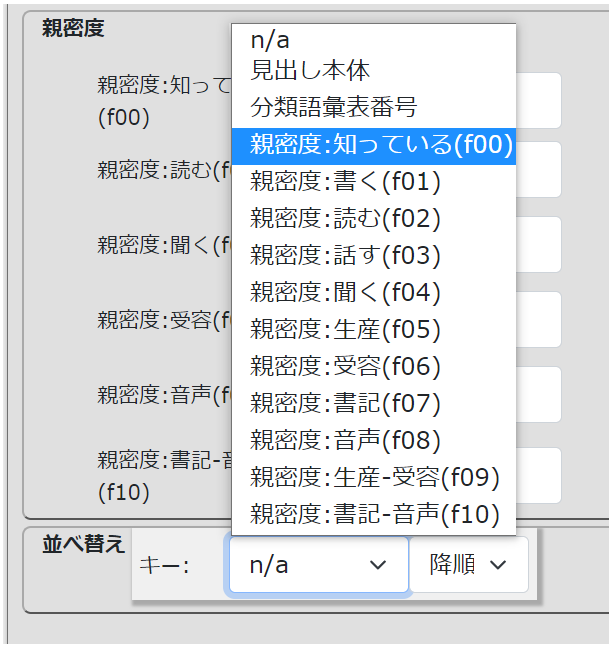

図 11　単語親密度による並び換えの指定

図 12　表示列幅の指定