クラウドソーシングによる単語親密度の推定

浅原 正幸 *
国立国語研究所

1. はじめに

辞書・語彙表は、語釈・音韻情報・形態論情報・品詞・統語情報・意味情報・用例といった言語学的知識の様々な情報を付与して編纂される。他の情報として言語運用者の主観的な情報などの付与が考えられる。NTT データベースシリーズ「日本語の語彙特性」の単語親密度はそのようなデータの一つである (天野・近藤 1999)。しかしながら、本調査は 20 年ほど前の調査であり、現在の言語運用を適切にモデル化できているとはいえない。

本研究では『分類語彙表』(国立国語研究所 2004) の見出し語に対して単語親密度を付与したので報告する。『分類語彙表』に対する単語親密度情報を Yahoo! クラウドソーシングを用いて収集する。3,392 人からなる実験協力者は、当該語を「知っている」「書く」「読む」「話す」「聞く」の 5 つの観点について、内省に基づき 1-5 の尺度をアンケート形式で付与する。各単語は少なくとも 16 人の尺度情報を収集した。これらの結果をもとにベイジアン線形混合モデル (Sorensen et al. 2016) を用いて、単語親密度を推定した。

本研究の貢献は以下のとおりである：

- クラウドソーシングに基づく単語親密度評定手法を提案した
- 単語親密度においては、知っているか否かだけでなく、{ 音声・書記 } と { 生産・受容 } の 2 軸の評価を行った
- 大規模な語彙に対する属性推定に対して、ベイジアン線形混合モデルを導入した

なお、本研究は 2017 年に相の類で事前調査した研究 (浅原 2017) を『分類語彙表』の区切り文字を除く全要素 100,830 語に拡張したものである。

2. 単語親密度情報の収集方法

『分類語彙表』に対する単語親密度情報を Yahoo! クラウドソーシングを用いて収集する。3,392 人からなる調査協力者は、当該語を「知る」「書く」「読む」「話す」「聞く」の 5 つの観点について、内省に基づき 1-5 の尺度をアンケート形式で付与する。各単語は少なくとも 16 人の尺度情報を収集した。

分類語彙表の見出し語 100,830 語を対象として、以下の 5 つの観点について Yahoo! クラウドソーシングによりアンケート調査を行った。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。

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) ー (5) よく知っている
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) ー (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) ー (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) ー (5) よく出現する

* masayu-a@ninjal.ac.jp

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) ー (5) よく出現する

「KNOW: 知っている」の観点で、その単語を知っているかどうかを確認するほか、書記言語か音声言語か（ {WRITE, READ} 対 {SPEAK, LISTEN}）、生産過程か受容過程か（ {WRITE, SPEAK} 対 {READ, LISTEN}）の 2 軸による 4 つの観点を確認した。データポイント数は 1,617,184 である。

3. 収集データの基礎統計

本節では収集したデータの基礎統計について示す。

表 1　単語親密度の基礎統計

順位	全評定値	KNOW	WRITE	SPEAK	READ	LISTEN
平均	2.818	3.814	2.390	2.511	2.734	2.639
標準偏差	0.962	0.931	0.696	0.824	0.766	0.837

表 1 に単語親密度の基礎統計を示す。知っているか否か (KNOW) については、3.814 と高い平均値を示しているが、それ以外の評定値については、2.390-2.734 と低い平均値を示している。標準偏差は各評定値 0.696-0.931 を示している。

表 2　単語親密度上位語

順位	全評定値（平均）	WRITE	SPEAK	READ	LISTEN
1	月曜	仕事する／をする	日曜日	利用	こんにちは
2	仕事する／をする	か月・箇月（かげつ）	でも	人（ひと）	昼御飯
3	利用	日曜日	昼御飯	でも	月曜
4	日曜日	話す	ティッシュ	月曜	寒い
5	か月・箇月（かげつ）	利用	利用	寒い	仕事する／をする

表 2 に単語親密度上位語を示す。KNOW は評定値の平均が 5 であるものが 250 語存在するので省略する。曜日や「仕事する」「利用」「でも」「寒い」などが上位語であった。なお、単語親密度最下位の語は、「宸翰」「那辺」「参籠」「スフ」「うずみひ」であった（全評定値が 1）。

表 3　書記言語 (WRITE, READ) －音声言語 (SPEAK, LISTEN)

順位	正方向（書記言語寄り）		負方向（音声言語寄り）	
1	上記	3.5125	レジ袋	-3.2500
2	前述する	2.5625	先っちょ	-2.8125
3	後述	2.5000	バイバイ	-2.7500
4	記	2.4735	ちょろまかす	-2.7500
5	在中	2.3125	ヨーグルト	-2.6875

次に書記言語寄りであるか音声言語寄りであるかを評価するために WRITE, READ の和から SPEAK, LISTEN の和を引いたものを評価する。表 3 に正方向上位語と負方向上位語を示す。正方向上位語には書面で利用される「上記」「前述する」「後述」「記」「在中」などが出現した。負方向上位語には音声言語寄りの言葉だけでなく、生活に身近な言葉「レジ袋」「ヨーグルト」などが出現した。

表 4　生産過程 (WRITE, SPEAK) −受容過程 (READ, LISTEN)

順位	正方向（生産過程寄り）		負方向（受容過程寄り）	
1	歌聖	0.8125	送検する	-3.1250
2	毛管	0.8125	書類送検	-2.8750
3	だるい	0.7500	殺害（さつがい・せつがい）	-2.6875
4	絆創膏	0.7500	西郷隆盛	-2.6875
5	上辺（うわべ）	0.7500	ＩＡＥＡ	-2.6875

最後に生産過程寄りか受容過程寄りかを評価するために、WRITE, SPEAK の和から READ, LISTEN の和を引いたものを評価する。表 4 に正方向上位語と負方向上位語を示す。正方向上位語には、専門的な用語「歌聖」「毛管」「絆創膏」など実験協力者の方が『自分は使うが一般の人は使わない』と判断した語が出現したほか、「だるい」など口にしやすいネガティブな語が出現した。一方、負方向上位語には、「送検する」「書類送検」「殺害（さつがい・せつがい）」など口にしにくいネガティブな語が出現したほか、「西郷隆盛」「ＩＡＥＡ」など、テレビ放送や新聞などで目にする語が出現した。

4. データの統計処理

本節では、実験協力者のバイアスを軽減するためのデータの統計処理方法について述べる。

「日本語の語彙特性」の単語親密度は、音声刺激と書記刺激を実験協力者に実験室で呈示していた。このため実験協力者に丁寧な教示を行い、ある程度の統制を行っていた。本研究ではクラウドソーシングを用いているために実験協力者の統制が困難であり、実験協力者ごとのバイアスが生じる。このバイアスを軽減させるためににベイジアン線形混合モデル (Sorensen et al. 2016) を用いて回帰する。実験協力者のバイアスをランダムスロープとしてモデル化すると同時に、単語毎の評定値についてもランダムスロープとしてモデル化し、それを推定された単語親密度として用いる。

以下、具体的な手法について解説する。

N_{word} は調査する単語（と観点）の数 (= 100,830 × 5)、N_{subj} は調査協力者の数 (= 3,392)、$i : 1 \ldots N_{word}$ が単語に対するインデックスで、$j : 1 \ldots N_{subj}$ が調査協力者に対するインデックスである。$y^{(i)(j)}$ は単語親密度（KNOW, WRITE, READ, SPEAK, LISTEN）の数で、次の正規分布としてモデル化する：

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

ここで σ は標準偏差である。平均 $\mu^{(i)(j)}$ は、切片 α と調査協力者のランダムスロープ $\gamma^{(}_{word}i)$ と単語のランダムスロープ $\gamma^{(}_{subj}j)$ からなる次の線形式でモデル化する：

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

各ランダムスロープについても次式のように正規分布としてモデル化する（ここで μ_{subj}, σ_{subj}, μ_{word}, σ_{word} は、各ランダムスロープの平均と標準偏差）

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

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

$\gamma_{subj}^{(j)}$ により被験者のゆれを吸収・統制するとともに、$\gamma_{word}^{(i)}$ を単語親密度として利用する。
推定は R および Stan を用いた。

5. おわりに

本稿では、クラウドソーシングによる単語親密度の推定方法について解説した。分類語彙表に含まれるすべての語句について、最低 16 人以上の評定情報を KNOW, WRITE, READ, SPEAK, LISTEN の 5 つの観点により収集した。評定情報をもとに、ベイジアン線形混合モデルにより、実験協力者間のバイアスを吸収しながら、単語ごとの評定値を推定した。本データは `github.com/masayu-a/WLSP-familiarity` からダウンロードできる。締切の都合上、統計処理の効果について定量的に評価することができなかったが、今後評価を進める。

今回評価した語彙集合は分類語彙表に基づく統語・意味情報ラベルが付与されている。本研究により、統語・意味情報ラベルごとの親密度が評価でき、基本レベルカテゴリが推定できると考える。

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および科研費 JP17H00917, JP18H05521, 18K18519 によるものです。

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.

国立国語研究所 (2004).『分類語彙表–増補改訂版』　大日本印刷, 東京.

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

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 巻, pp. 45–50.