

国立国語研究所学術情報リポジトリ

分類語彙表に対する単語親密度情報付与

メタデータ	言語: Japanese 出版者: 公開日: 2021-03-09 キーワード (Ja): キーワード (En): 作成者: 浅原, 正幸 メールアドレス: 所属:
URL	https://doi.org/10.15084/00003202

『分類語彙表』の見出し語 95,071 語に対する単語親密度付与 知っている (KNOW), 書く (WRITE), 読む (READ), 話す (SPEAK), 聞く (LISTEN) の5観点を Yahoo クラウドソーシングにより収集し、ベイジアン線形混合モデルで推定

調査方法

分類語彙表 [国立国語研究所 2004]
語を意味によって分類・整理したシソーラス (類義語集)

分類番号の構造 例: よう (分類番号: 3.1300)

類	部門	中項目	分類項目
相 (3)	抽象的關係 (1)	様相 (13)	様相・情勢 (1300)

Yahoo! クラウドソーシングによる調査

設定した質問ID: 1A3991A079920

以下の単語についてお答えください

言い換す

単語の意味は知っていますか?

全く知らない あまり出ない
 どちらともいえない ほとんど知っている
 よく知っている

どのくらい頻りに出ているものに出会えますか?

全く出ない あまり出ない
 どちらともいえない たまに出る
 よく出る

どのくらい頻りに出ているものに出会いますか?

全く出ない あまり出ない
 どちらともいえない たまに出る
 よく出る

どのくらい頻りに出ているものに出会いますか?

全く出ない あまり出ない
 どちらともいえない たまに出る
 よく出る

【参考情報: 079920-用(活動)登録(任意・匿名)】

2018年11月に実施
3,392人の実験協力者から
1見出し語あたり16人以上のデータ
(5観点)を収集
データポイント数 1,617,184件

5段階評価

- 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/たまに出る
- 5 よく知っている/よく出現する

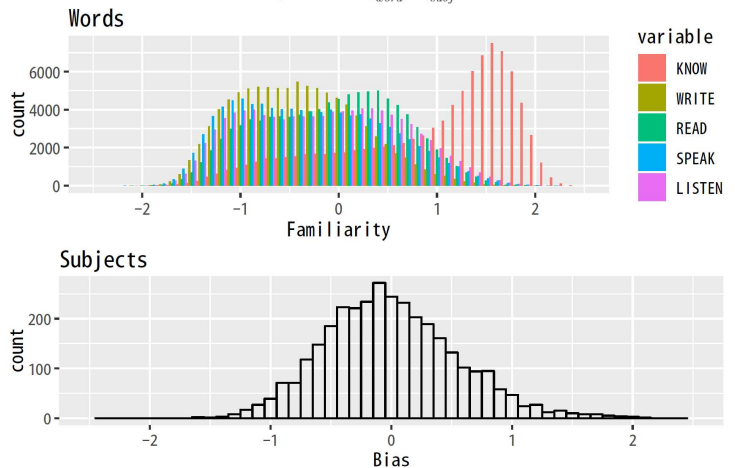
ベイジアン線形混合モデルによるモデル化

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), $i: 1 \dots N_{word}$ が単語に対するインデックスで, $j: 1 \dots N_{subj}$ 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) の値で, 次の正規分布としてモデル化する:

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

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:

$$\mu^{(i,j)} = \alpha + \gamma_{word}^{(i)} + \gamma_{subj}^{(j)}$$



推定結果 (各観点の上位 or 下位 10位まで)

知っている	知らない	書く	書かない	読む	読まない
全員	うずみひ	一月・箇月 (かげつ)	玉章 (たまざさ)	聞く	玉章 (たまざさ)
恋人	玉章 (たまざさ)	聞く	痛罵	休む	あおのけ
翌朝 (よくあさ)	御稜威 (みいつ)	仕事する/ををする	暴戻 (ぼうれい)	書く	鞆掌 (おうしょう) する
退社する	繞 (にょう)	話す	鞆掌 (おうしょう) する	登録する	撃攘
再会	鞆掌 (おうしょう) する	見る	縊死 (いし)	自分	席題
本社	スフ	自分	撃攘	見る	暖国
入社	馳名	力 (ちから)	席題	明日 (あした)	退京
人見知りする	笈指 (おいずり)	口	馳名	人 (ひと)	御稜威 (みいつ)
持ち帰る	宇内 (うだい)	作る	あおのけ	終わる	劫を経る
ストロー	賢察	食料品	みず垣	今度	馳名

生産上位	受容上位	話す	話さない	聞く	聞かない
聞く	聞く	聞く	みず垣	聞く	玉章 (たまざさ)
終わる	終わる	終わる	玉章 (たまざさ)	終わる	劫を経る
一月・箇月 (かげつ)	休む	野菜	痛罵	野菜	あおのけ
今度	書く	今度	劫を経る	食べ物	鞆掌 (おうしょう) する
今度	今度	母親	薫染	会う	微賤
休む	見る	食べ物	使臣	休む	少時
話す	作る	会う	吏道	書く	吏道
見る	明日 (あした)	年を取る	微賤	今度	放念
自分	食べ物	書く	馬手 (めて)	おはよう [~ございます]	席題
書く	今日 (きょう)	今度	地映	寝る	縊死 (いし)

書記上位	音声上位	生産-受容上位	受容-生産上位	書記-音声上位	音声-書記上位
聞く	聞く	毛管	送検する	上記	レジ袋
一月・箇月 (かげつ)	終わる	物心 (ぶっしん)	右翼	追伸	先っちょ
自分	野菜	消却する	書類送検	前述する	ちよろまかす
休む	今度	絆創膏	巡業する	後述	パイパイ
作る	食べ物	ふたとせ	西郷隆盛	記	ヨーグルト
話す	会う	揚げなべ	殺害 (さつがい・せつがい)	前略	ドライヤー
終わる	書く	吟詠する	革命児	在中	まんま [その~]
今度	今度	たるい	護衛する	アンバサンド [&]	それではまた
私 (わたし)	休む	上辺 (うわべ)	識者	句読点	鼻水
登録する	年を取る	幽寂	再審	下記	どっこいしょ