分類語彙表に対する単語親密度情報付与

分類語彙表に対する単語親密度情報付与

浅原正幸　コーパス開発センター

『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

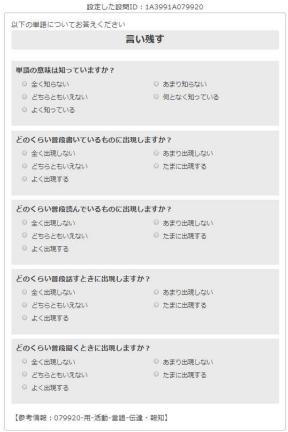
調査方法

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)	抽象的関係 (.1)	様相 (.13)	様相・情勢 (.1300)

Yahoo! クラウドソーシングによる調査



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/たまに出現する
5 よく知っている/よく出現する

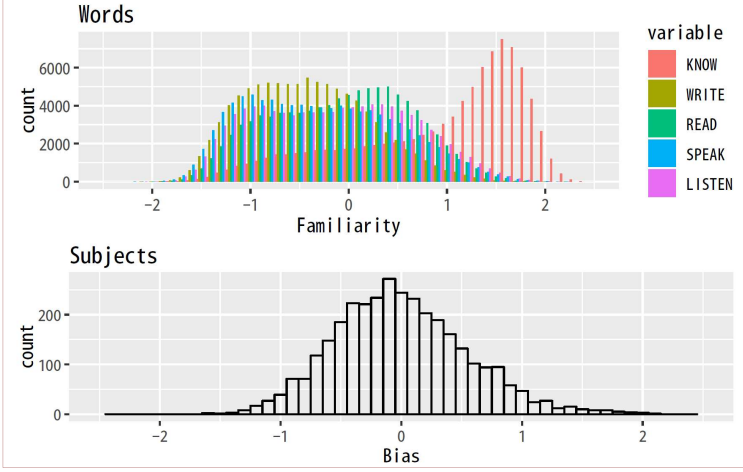
ベイジアン線形混合モデルによるモデル化

N_{word} は調査する単語（と観点）の数 $(= 100{,}830 \times 5)$、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）の値で、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 $\gamma^{(i)}_{word}$ と単語のランダムスロープ $\gamma^{(j)}_{subj}$ からなる次の線形式でモデル化する：

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



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
全員	うずみひ
恋人	玉章（たまずさ）
翌朝（よくあさ）	御稜威（みいつ）
退社する	繞（にょう）
再会	執掌（おうしょう）する
本社	スフ
入社	驍名
人見知りする	笈摺（おいずり）
持ち帰る	宇内（うだい）
ストロー	賢察

書く	書かない
-か月・箇月（かげつ）	玉章（たまずさ）
聞く	痛罵
仕事する／をする	暴戻（ぼうれい）
話す	執掌（おうしょう）する
見る	縊死（いし）
自分	撃攘
力（ちから）	席題
口	驍名
作る	あおのけ
食料品	みず垣

読む	読まない
聞く	玉章（たまずさ）
休む	あおのけ
書く	執掌（おうしょう）する
登録する	撃攘
自分	席題
見る	暖国
明日（あした）	退京
人（ひと）	御稜威（みいつ）
終わる	劫を経る
今度	驍名

生産上位	受容上位
聞く	聞く
終わる	終わる
-か月・箇月（かげつ）	休む
今度	書く
今度	今度
休む	見る
話す	作る
見る	明日（あした）
自分	食べ物
書く	今日（きょう）

話す	話さない
聞く	みず垣
終わる	玉章（たまずさ）
野菜	痛罵
今度	劫を経る
母親	薫染
食べ物	使臣
会う	吏道
年を取る	微賤
書く	馬手（めて）
今度	地峡

聞く	聞かない
聞く	玉章（たまずさ）
終わる	劫を経る
野菜	あおのけ
食べ物	執掌（おうしょう）する
会う	微賤
休む	少時
書く	吏道
今度	放念
おはよう［～ございます］	席題
寝る	縊死（いし）

書記上位	音声上位
聞く	聞く
-か月・箇月（かげつ）	終わる
自分	野菜
休む	今度
作る	食べ物
話す	会う
終わる	書く
今度	今度
私（わたし）	休む
登録する	年を取る

生産-受容上位	受容-生産上位
毛管	送検する
物心（ぶっしん）	右翼
消却する	書類送検
絆創膏	巡業する
ふたとせ	西郷隆盛
揚げなべ	殺害（さつがい・せつがい）
吟詠する	革命児
だるい	護衛する
上辺（うわべ）	識者
幽寂	再審

書記-音声上位	音声-書記上位
上記	レジ袋
追伸	先っちょ
前述する	ちょろまかす
後述	バイバイ
記	ヨーグルト
前略	ドライヤー
在中	まんま［その～］
アンパサンド［&］	それではまた
句読点	鼻水
下記	どっこいしょ

浅原正幸（2019）「クラウドソーシングによる単語親密度の推定」言語処理学会第25回年次大会発表論文集 pp.45-48.【NLP2019言語資源賞受賞】
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17H00917, 18H05521,18K18519, 19K00591, 19K00655 によるものです。