クラウドソーシング結果の可視化手法と統計処理

人間文化研究機構 国立国語研究所　　浅原　正幸

1　はじめに

言語使用の実態調査手法の一つに社会調査がある。従来は面接調査・郵送調査・電話調査など高コスト（時間・費用）であったが、近年クラウドソーシングサイトのマイクロタスクに基づく電子調査により、1 件あたり数円の低コストで数千人分のデータが 1 日で集めることが可能になりつつある。

本発表では、クラウドソーシング結果の可視化手法と統計処理手法についていくつか紹介する。ここで扱うデータは、事例を見せたうえで実験協力者の評定値をサーストン法（等現間隔法）で得た離散値の 1 次元データを扱う。事例として、以下の 3 つを紹介する：

- 代表値の可視化
  指標比喩の評定データ (Kikuchi et al., 2019)
- ネットワークの可視化
  意味の類似度の評定データ (西内, 2019; 加藤, 2019)
- 実験協力者間の揺れの統制
  単語親密度の評定データ (浅原, 2019)

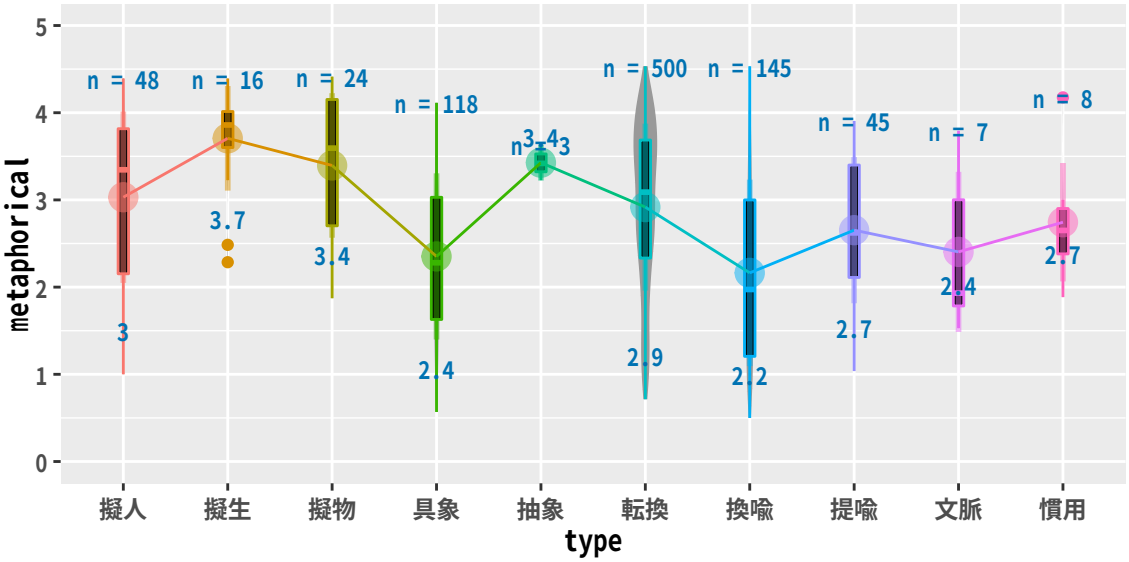

Figure 1: 代表値の可視化：指標比喩データベースの種別ごとの比喩性判定

2　可視化

2.1　代表値の可視化：指標比喩の比喩性判定

まず代表値の可視化について簡単に触れる。以下では指標比喩データベースに対する実験協力者評定値 Kikuchi et al. (2019) を用いて説明する。本データは『現代日本語書き言葉均衡コーパス』から言語学の専門家により人手で抽出した指標比喩用例 816 事例を、擬人・擬生・擬物・具象・抽象・その他転換・換喩・提喩・文脈比喩・慣用表現などの種別を付与したうえで、実験協力者から「比喩表現を含んでいるか」（全く違う） 0-5 （そう思う）の 6 段階評価で回答を得た。異なり 1657 人の実験協力者から、1 事例あたり 22-66 人（平均 33 人）の回答を得、「事例単位の平均値」をとったものを可視化する。

Figure 1 に「事例単位の平均値」の種別ごとの代表値を可視化したものを示す。このグラフでは横軸に種別を示し、縦軸に比喩性（0-5 の値）を示す。グラフでは、箱ひげ図・バイオリン図・平均値の折れ線グラフを重ね合わせている。箱ひげ図により最小値・第 1 四分位・中央値・第 3 四分位・最大値と外れ値を表す。バイオリン図により確率密度関数の推定結果（カーネル密度推定）を表す。各図の上部に度数を表し、下部と折れ線により平均値を表す。このようにして要約統計量を可視化することができる。

2.2　ネットワークの可視化：意味の類似度判定

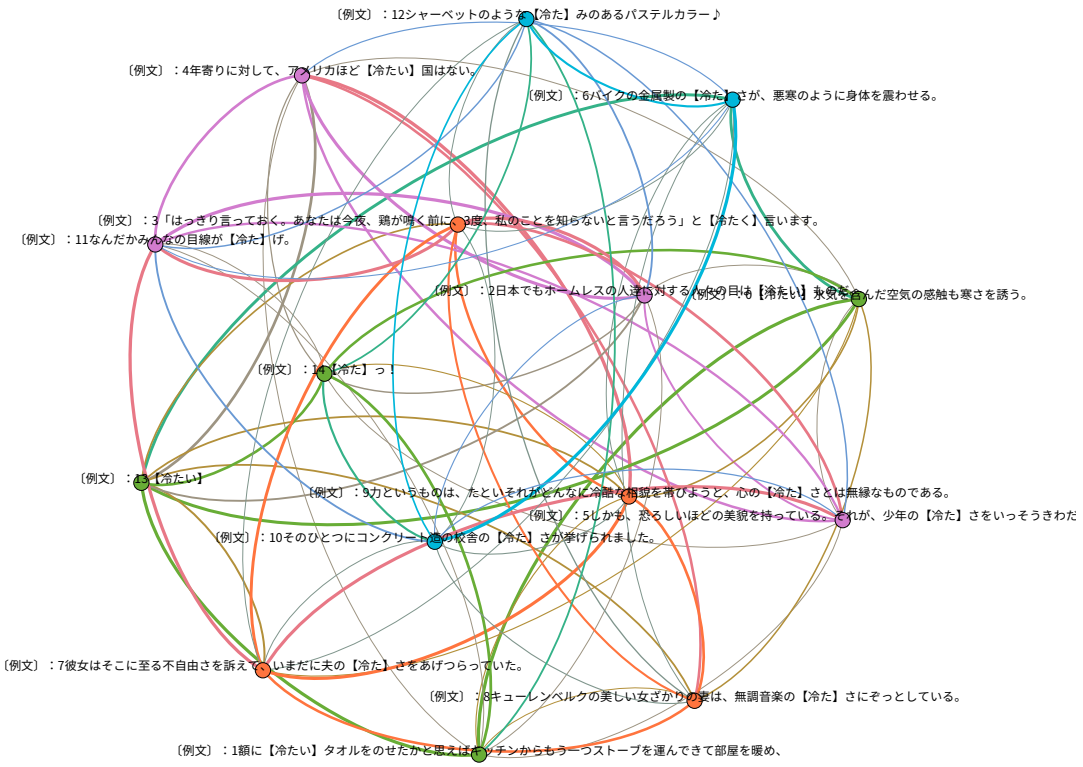

Figure 2: ネットワークの可視化：「冷たい」の意味の類似度判定

次にネットワークの可視化について説明する。

最初の例では複数の例文間の類似度判定を行ったものの可視化について紹介する。「冷たい」を含む 15 例文について、異なり 320 人の実験協力者から、語の意味の類似度判定を （全く違う）0-5（全く同じ）の 6 段階評価で回答（1 事例あたり 20-40 人）を得た。例文を先に見せたか後に見せたかで個別に集計するために、15 × 14 の順列を可視化する必要がある。Figure 2 に Gephi [1] を用いて可視化したものを示す。頂点が例文を表し、辺が 2 つの例文の類似度を表す。頂点の配置は Fruchterman-Reingold アルゴリズムによる。このアルゴリズムにおいて各頂点は、非連結の他の頂点から斥力を、連結の他の頂

[1] https://gephi.org/

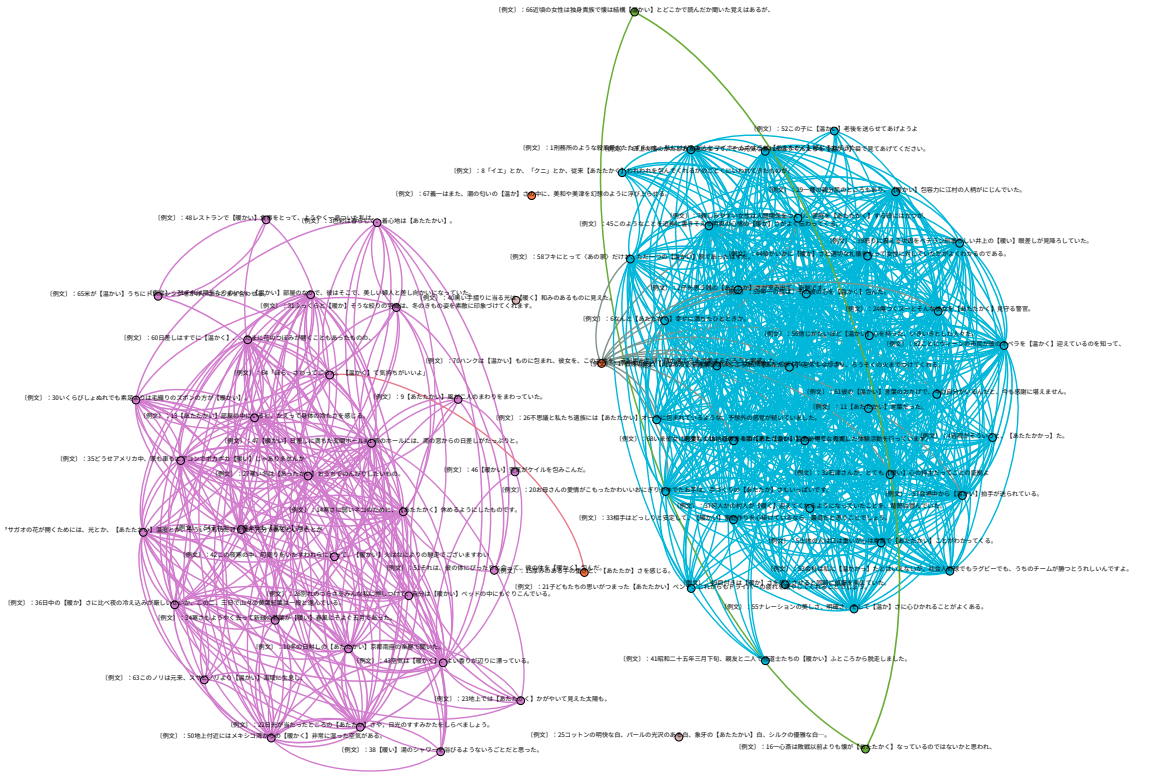

Figure 3: ネットワークの可視化：「暖かい」の意味の類似度判定 (類似度平均 3.5 以上)

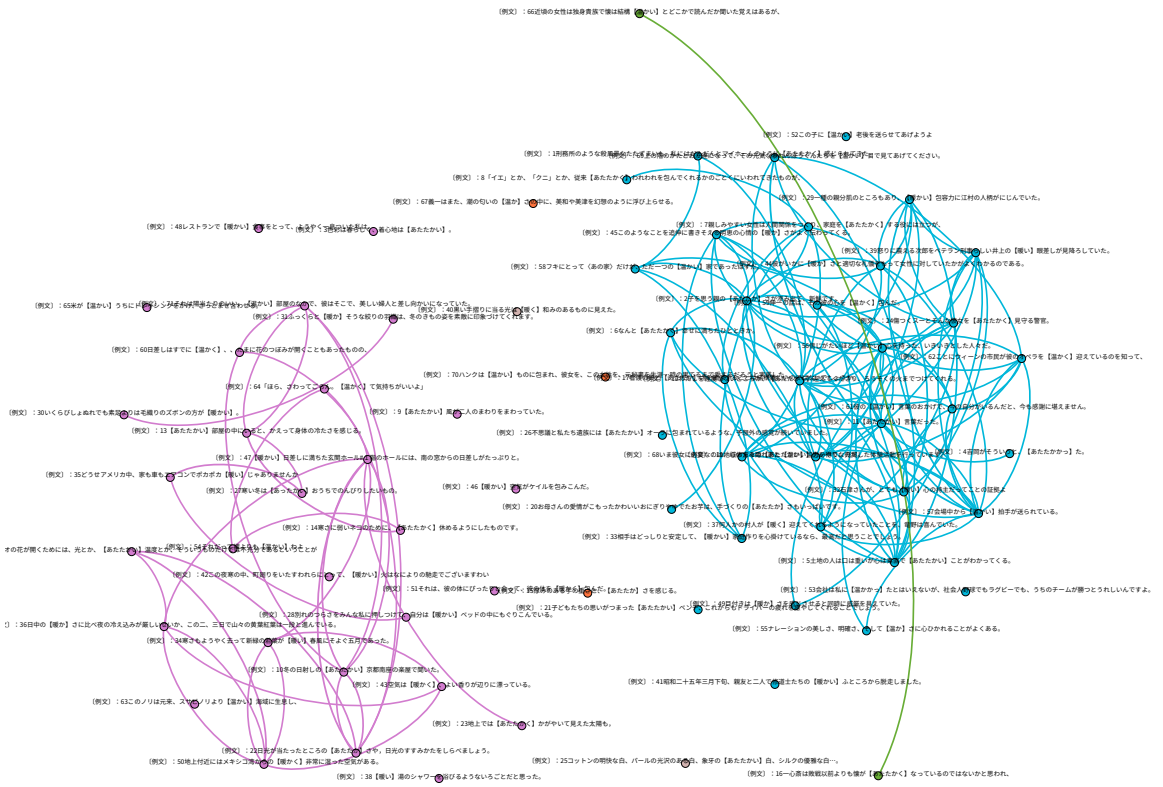

Figure 4: ネットワークの可視化：「暖かい」の意味の類似度判定 (類似度平均 4.0 以上)

点から引力を受けるように配置される。本調査ではすべての順列について評価されるため完全有向グラフであるために幾何学的な模様になる。類似度は頂点の太さによってあらわされているが、頂点の配置に影響を与えない。このためわかりやすくするためにコミュニティ検出機能を用いて色分けをおこなっている。Louvain 法に基づく、辺媒介中心性に基づくコミュニティ検出結果から、紫・青・緑・橙の 4 色に辺と頂点を塗り分けている。

次の例も同様に複数の例文間の類似度判定を行ったものである。「あたたかい」を含む 71 例文について同様の調査を行った（実験協力者異なり 8530 人，1 対あたり 50 人）。頂点の配置は、辺の重みに基づく力指向アルゴリズムにして配置した。全連結すると 71×70 の順列の辺を可視化する必要がある。これを回避するために Figure 3 では類似度が平均 3.5 以上の辺のみを示し、Figure 4 では類似度が平均 4.0 以上の辺のみを示した。

このようにして、事例間の類似度の分布を俯瞰することができる。

3 統計処理

本節では，クラウドソーシングデータの統計処理手法について述べる。観測データは実験協力者間のばらつきを含んでおり，これを考慮する必要がある。その取扱手法について示す。

3.1 統計処理の概要

まず、個体差（ばらつき）を考慮する最初のステップとして分散分析について説明する。分散分析においては、観測データの誤差変動と要因、さらにそれらの交互作用に基づく統計的仮説検定を行う。分散成分の平方和のを分解したうえで、誤差による変動から注目する（固定）要因による変動を分離し、誤差と要因を分離するということが行う。しかしながら、naive な手法のため、多数の要因を入れる場合に多重比較の問題が発生する。このうち残差が多変量正規分布に従うものが一般線形モデル (LM) であり、任意の分布としたものが一般化線形モデル (GLM) と呼ぶ。

GLM は、注目している固定要因のみを分析する枠組だが、混合モデルは個体差（実験協力者差・用例差）をランダム要因として考慮する。これを一般化線形混合モデル (GLMM) と呼ぶ。但し、あくまで最尤推定をおこなうだけなので、多変数の場合、収束しなかったり、あてはまりがわるかったりする。なお、この最尤推定による手法をベイズ主義的な手法に対して、頻度主義的な手法と呼ぶ。結果の判定には帰無仮説を導入し、いわゆる有意差をみて判定する。有意差は帰無仮説と対立仮設の間でどちらを棄却するのかについて判定を行う枠組だが、多変数の場合、仮説の組合せを検討する必要があり、多重比較の問題がまた発生する。

ベイジアン線形混合モデル (BLMM) は、階層ベイズモデルで固定要因とランダム要因を扱う線形モデルである。頻度主義的な最尤推定ではなく、ベイズ主義的な MCMC サンプリングを行い、事後平均を推定する。GLMM と同様に、ランダム要因により個体差を考慮することができる。BLMM の簡単なチュートリアルとして Sorensen et al. (2016) がある。

3.2 統計処理事例：単語親密度の推定

分類語彙表の見出し語 100,830 語を対象として、単語親密度情報を Yahoo! クラウドソーシングを用いて収集した。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。「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。各単語は少なくとも 16 人の尺度情報を収集した。データポイント数は 1,617,184 であった。

本データ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 BLMM 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：

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

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で、0 と 1 に設定）

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

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

$\gamma^{(j)}_{subj}$ により実験協力者のバイアスを吸収・統制するとともに、$\gamma^{(i)}_{word}$ 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。コードを 7 に示す。学習は warm up 100 iterations のあと 5000 iterations を 4 chains 実施し、すべてのモデルは収束した。

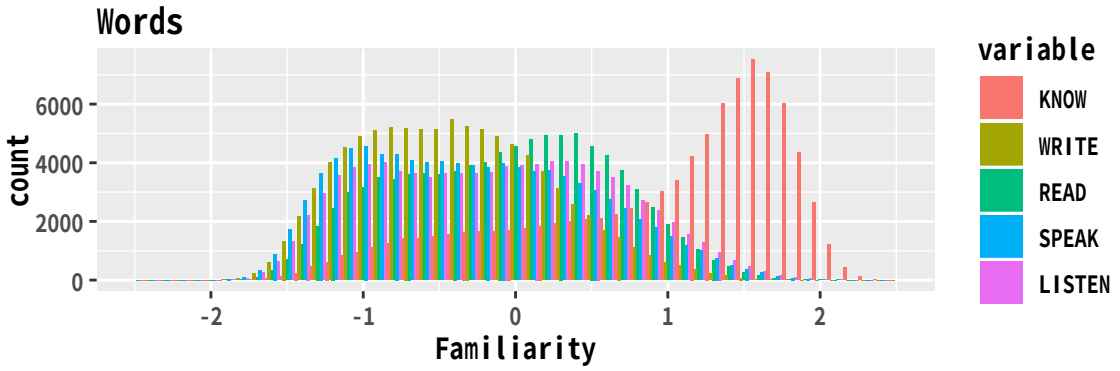

Figure 5: 単語親密度のモデル化（$\gamma^{(i)}_{word}$）: 5 観点の分布

Figure 5 に推定された単語親密度の結果を図示する。知ってはいるけれど (KNOW)、その使用実態 (WRITE, READ, SPEAK, LISTEN) については低い傾向がみられる。書記 (WRITE, READ) のほうが音声 (SPEAK, LISTEN) よりも優勢で、受容 (READ, LISTEN) のほうが生産 (WRITE, SPEAK) よりも優勢であることがモデル化できている。

Figure 6 に実験協力者のバイアスの事後平均の分布を示す。実験協力者の語彙数に基づくバイアスを正規分布でモデル化しており、単語親密度の推定にはこのバイアスを減じて評価している。

4 おわりに

本発表では、クラウドソーシング結果の可視化手法と統計処理手法について解説した。まず、指標比喩の評定データの要約統計量の可視化手法と 2 つの例文中に出現する語義の類似度評価の可視化手法を示した。次に、実験協力者のバイアスを吸収するための統計処理事例として、単語親密度の推定手法について解説した。

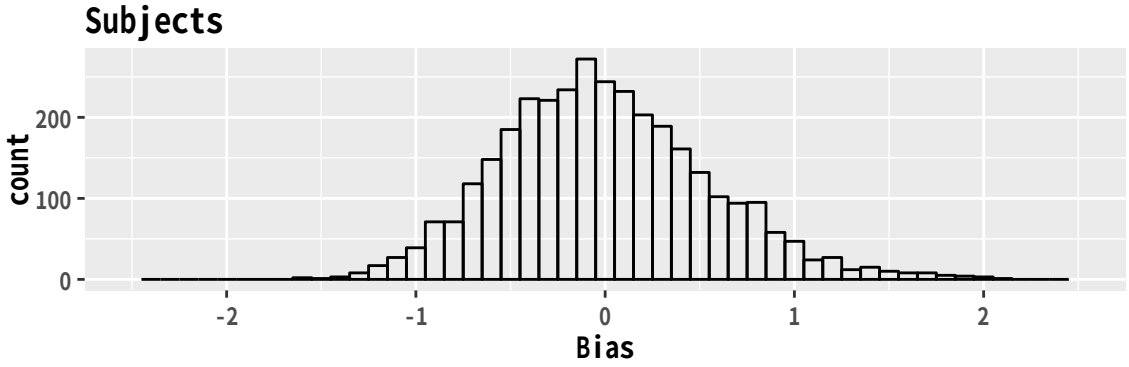

Figure 6: 単語親密度のモデル化（$\gamma_{subj}^{(j)}$）：実験協力者のバイアスの分布

```
data {
        int<lower=0> N;           // number of data points
        int<lower=0> N_word;      // number of words
        int<lower=0> N_subj;      // number of items
        int<lower=0,upper=N_word> word[N];  // word id
        int<lower=0,upper=N_subj> subj[N];  // subject id
        int<lower=1, upper=5> y[N]; // rating
    }
parameters {
        real alpha;      // intercept
        vector[N_word] gamma_word; // word slope
        vector[N_subj] gamma_subj; // subj slope
}
model {
        real mu;
        gamma_word ~ normal(0,1); // prior for word
        gamma_subj ~ normal(0,1); // prior for subject
        for (i in 1:N) {
            mu = alpha + gamma_word[word[i]] + gamma_subj[subj[i]];
            y[i] ~ normal(mu,1);
        }
}
```

Figure 7: stan のコード例

Acknowledgement

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・所長裁量経費プロジェクト 2018 および科研費 17H00917, 18H05521, 18K18519, 19K00591, 19K00655 によるものです。

References

Kikuchi, Rei, Sachi Kato, and Masayuki Asahara (2019) “Collecting Figurative Expressions Using Indicators and a Sementic Tagged Japanese Corpus,” in *the International Cognitive Linguistic Conference (ICLC15)*.

Sorensen, Tanner, Sven Hohenstein, and Shravan Vasishth (2016) “Bayesian linear mixed models using Stan: A tutorial for psychologists, linguists, and cognitive scientists,” *Quantitative Methods for Psychology*, Vol. 12, pp. 175-200.

加藤祥 (2019) 「クラウドソーシングによる語義調査」,『日本言語学会第 158 回大会』.

西内沙恵 (2019) 「クラウドソーシングによる述定・装定の用法分析」,『日本言語学会第 158 回大会』.

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』, 45-48 頁.