『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

加藤 祥
目白大学

浅原 正幸
国立国語研究所

masayu-a@ninjal.ac.jp

概要

本研究では『現代日本語書き言葉均衡コーパス』の短単位動詞・長単位自立語・文節を刺激として、一般の方の印象評定情報を収集した。「自然さ」「わかりやすさ」「古さ」「新しさ」「比喩性」の 5 観点について「0:まったく違う」から「5:そう思う」の 6 段階評価を行った。これらの印象評定情報に基づいて、既存の代表義情報を線形回帰するモデルを構築し、コーパスにあてはめることにより典型用例を抽出することを試みた。

1 はじめに

本研究では、クラウドソーシングにより、印象評定情報を『現代日本語書き言葉均衡コーパス』に付与したので報告する。

『分類語彙表』に対する単語親密度データベース WLSP-Familiarity [1, 2] [1)]は辞書の見出しを刺激として、「知っている」「書く」「読む」「話す」「聞く」の 5 観点について評定値を収集したものである。一方、単語親密度は単語に対する評定値であり、実際に使用されている文脈でどのように捉えられているかがわからないという問題があった。さらに多義語については語義ごとの親密度の判定が困難であった。

そこで『現代日本語書き言葉均衡コーパス』[3]（以下 BCCWJ）の文脈を呈示し、その印象評定情報を収集したので報告する。具体的には、国語研短単位動詞・国語研長単位自立語・文節単位に「自然さ」「わかりやすさ」「古さ」「新しさ」「比喩性」の 5 観点について 0(全く違う)〜5(そう思う) の 6 段階評価を収集した。本稿では、印象評定情報の収集方法を解説するとともに、データの基礎統計などについて示す。さらに、代表義情報をこれらの印象評定情報に基づき回帰することで、コーパス中の典型用例を抽出することを試みたので報告する。

1） https://github.com/masayu-a/WLSP-familiarity

2 関連研究

2.1 印象評定情報

NTT データベースシリーズ『日本語の語彙特性』[2)]は、人間の言語機能の解明を目指し、様々な観点で語彙の特徴を検証した世界最大規模のデータベースである。『日本語の語彙特性』の中には、そのほかに単語親密度・単語表記の妥当性・単語アクセントの妥当性・漢字親密度・漢字複雑度・漢字の読みの妥当性・単語心像性などの人の主観的な評定データのほか、新聞などの語彙の出現頻度に基づく客観的なデータも収録されている。その中で 単語親密度データベース（平成版）[4, 5] は語彙の親密度（なじみの有無）を収集した先進的な語彙データベースである。また、最初の調査から年月を経て、人の語彙のとらえ方が変遷しつつあるなか、単語親密度データベース（令和版）[6] も構築され、世界最大規模のデータベースが公開された [7] [3)]。さらに、単語心像性データベース [8] では、文字刺激・音声刺激に対する「意味内容を感覚的にイメージする際の容易さ」の収集が行われた。

国語研では、『分類語彙表』に対する単語親密度の推定 [1, 2] を継続的に進めているほか、種々の語彙表を公開している [9, 10][4)]。これらの国語研の語彙データベースは、多義語について人がどのように語彙をとらえているかについては明らかにできないものであった。多義語の語義調査を目指し、2021 年に語義情報を付与されている IPAL 辞書の用例に対して印象評定情報付与 [11, 12] を試験的に行った。本研究では、同研究を BCCWJ に拡張し、日本語の多義語の印象評定情報の付与を進める。

2） https://www.sanseido-publ.co.jp/publ/ep/RD/RD04.html
3） https://nttnttprint.com/lexicon-db/
4） http://doi.org/10.15084/00003472, http://doi.org/10.15084/00003473

2.2 中心義・基本義・代表義・典型用例

多義語における語義の認定は、当該言語において定着した意味として、一般的に辞書へ記載されるものと考えられる。瀬戸 [13]、多義の記述において、中心義の設定、意義の認定、意義関係の明示、意義の配列を議論する。多義語の語義認定には、関連語の対応による認定 [14, 15]、個別義認定の分離・統合テスト [16] などの認定基準が検討されている。

多義語には何らか中心的な意味が存在する [17] とされ、中心義・基本義・代表義などと呼ばれる。派生関係を考える場合には歴史的な変化に基づき出現年代順が重要視されるが、多義構造は再編成されるために、現在一般の方が想定する典型性のある中心的な語義が必ずしも歴史的な変化に基づくとは限らない。中心的な意味は一様に定められるものではなく、例えば瀬戸 [13] は (i) 文字通り、(ii) 他意義の前提、(iii) 具象性が高い、(iv) 認知しやすい、(v) 想起しやすい、(vi) 用法上の制約を受けにくい、(vii) 意義展開の起点（接点）となることが多い、(viii) 言語習得の早い段階で獲得される、(ix) 使用頻度が高いの 9 特性を挙げている。

言語資源としては、『分類語彙表』増補改訂版の多義語に対して、専門家により代表義を付与する作業を実施したものがある [18]。BCCWJ に付与された分類語彙表番号データ [19] からコーパス中の専門家による代表義認定用例が認定できるほか、IPAL 辞書における語義に対しても分類語彙表番号が付与されていることから、IPAL 辞書における代表義を認定することも可能である。しかしながら、これらの認定は専門家によるものであり、一般の読み手の判断とは差のある可能性も考えられる。先に述べたとおり、 IPAL 辞書の用例に対して印象評定情報付与 [11] を行っていることから、専門家が代表義と認定する用例が、一般の読み手がどう捉えるのかについて線形回帰により求めることも可能になった [12]。本研究では、この線形回帰式を今回収集した BCCWJ に付与された印象評定情報にあてはめることで、代表義用例および典型用例の抽出を試みる。

3 データの収集手法

本節ではデータの収集方法について示す。

収集対象は BCCWJ のうち、短単位に基づいて語義情報が付与されている BCCWJ-WLSP (書籍, 新聞, 雑誌) [19] と読み時間情報が付与されている BCCWJ-SPR (書籍, 教科書) [20] とする。

前者の BCCWJ-WLSP は、BCCWJ コアデータの書籍, 新聞, 雑誌サンプルの一部（約 33 万語）の短単位自立語に対して分類語彙表に基づく語義を付与したものである。分類語彙表に基づく語義情報と評定情報を対照することを目的として、この BCCWJ-WLSP のサンプルに対して、試験的に短単位動詞と短単位サ変名詞+「為る」を対象として、1 例あたり 20 人の評定値を 2021 年 4 月 5 日〜5 月 3 日に収集した。さらに長単位自立語に対して、1 例あたり 10 人の評定値[5] を 2021 年 11 月 17 日〜12 月 6 日に収集した。

後者の BCCWJ-SPR2 は、BCCWJ のコアデータの書籍と国語科教科書サンプルを、文節単位の自己ペース読文法により、読み時間データを収集したものである。読み時間のふるまいを評定情報で説明することを目的として、この BCCWJ-SPR2 のサンプルに対して、文節を対象として、1 例あたり 10 人の評定値を 2021 年 11 月 17 日〜12 月 6 日に収集した。

図 4 印象評定情報収集画面

評定情報収集画面を図 4 に示す。画面の上に 1 文単位で例文を表示し、判定してもらいたい表現を【】

5) 分散が少ない実験協力者の回答が多いサンプルを中心に 10 人分の評定値を追加で収集した。

収集対象	収集単位	表現数	データポイント数	収集時期
BCCWJ-WLSP (書籍, 新聞, 雑誌)	短単位動詞	38,004	764,700	2021 年 4 月 5 日〜5 月 3 日
BCCWJ-WLSP (書籍, 新聞, 雑誌)	長単位自立語	122,173	1,227,060	2021 年 11 月 17 日〜12 月 6 日
BCCWJ-SPR2 (書籍, 教科書)	文節	135,342	1,358,650	2021 年 11 月 17 日〜12 月 6 日

表 1　評定値の収集対象

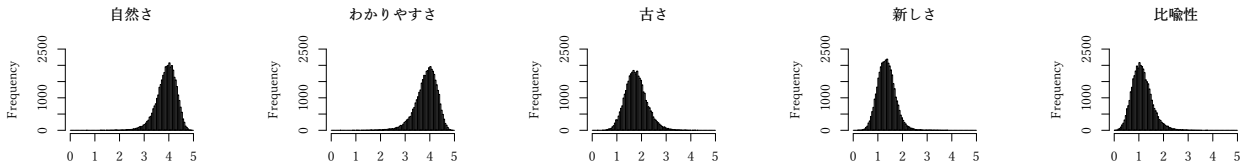

図 1　評定値の分布（BCCWJ-WLSP：短単位動詞：ビン幅 0.05）

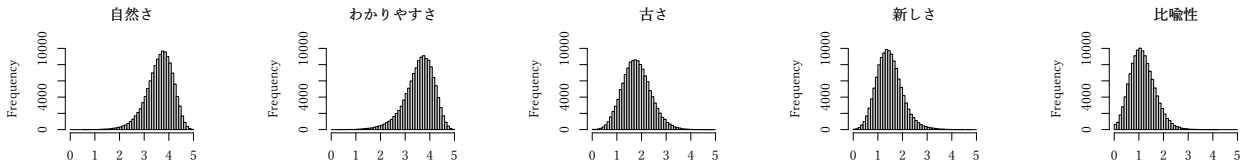

図 2　評定値の分布（BCCWJ-WLSP：長単位自立語：ビン幅 0.10）

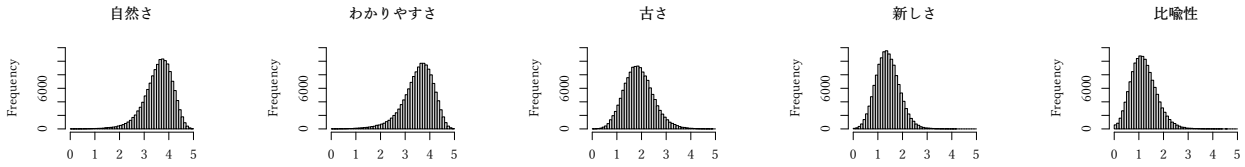

図 3　評定値の分布（BCCWJ-SPR2：文節：ビン幅 0.10）

で示した。評定値は、「自然さ」「わかりやすさ」「古さ」「新しさ」「比喩性」の 5 観点について、「0: まったく違う」から「5:そう思う」までの 6 段階評価とした。実験協力者は日本の Yahoo! クラウドソーシングのアカウントを持つ 20 歳以上の方とし、1 回答あたり 1 円相当の謝礼ポイントを支払った。また、90%以上について同一の回答をした方を、回答できないような処理を随時行った。

4　データの基礎統計

図 1, 2, 3 に、BCCWJ-WLSP の動詞短単位・BCCWJ-WLSP の長単位自立語・BCCWJ-SPR2 の文節に対する表現ごとの平均評定値のヒストグラムを示す。いずれのサンプルも書籍・新聞・雑誌・教科書などで公刊されているものなので、自然さとわかりやすさが高い。また書籍・新聞・雑誌については 2001-2005 年に公刊されたものを対象としているが、全体的に古くも新しくもない傾向が見られた。さらに比喩性についても低い傾向が見られた。

付録の 4 に各評定値の上位・下位事例を示す。

5　評定情報に基づく典型用例の推定

瀬戸 [13] の中心義特性	自	わ	古	新	比
(i) 文字通り			+		
(ii) 他意義の前提				-	
(iii) 具象性が高い		+			
(iv) 認知しやすい		+			
(v) 想起しやすい		+			
(vi) 用法上の制約を受けにくい	+				
(vii) 意義展開の起点（接点）となることが最も多い			+		-
(viii) 言語習得の早い段階で獲得される		+			
(ix) 使用頻度が高い				-	

表 2　瀬戸の中心義特性と本研究の印象評定情報との関係

同様の評定情報をすでに IPAL 辞書用例に付与しており [11]、この 5 つ組の評定情報から代表義性を線形回帰する試みが行われている [12]。これは、瀬戸 [13] の中心義特性を印象評定情報として定義しな

おし (表 2)、一般の方の印象評定の組み合わせから、基本義・代表義・典型用例を推定する試みである。

同研究では、代表義情報を 5〜1 の代表義度[6] として数値化したものを、印象評定値を固定効果とし、用例をランダム効果とした次式による一般化線形混合モデルにより回帰して行った。

代表義度 ~ 自然さ+わかりやすさ+古さ+新しさ+比喩性 + (1| 用例)

評定値	固定効果推定値	(標準誤差)
自然さ	+0.012	(0.008)
わかりやすさ	*** +0.033	(0.008)
古さ	*** -0.015	(0.004)
新しさ	*** +0.018	(0.004)
比喩性	*** -0.024	(0.004)
切片	*** +1.965	(0.071)
データポイント数		56,120

表 3　IPAL 用例における代表義度の回帰（動詞） [12]

推定した固定効果推定値を表 3 に示す。表 2 の瀬戸の中心義特性との関係では、古さが+((i) 文字通り、(vii) 意義展開の起点（接点）となることが最も多い)、新しさが-((ii) 他意義の前提、(ix) 使用頻度が高い) を想定していたが、得られた推定値は古さが-、新しさが+の係数であった。

以下では短単位動詞の評定結果に基づき、より代表義性の高い「典型用例」を抽出することを試みる。具体的には、同研究で得られた次の線形回帰式をあてはめる：

推定代表義度 (動詞) := 0.012 × 自然さ + 0.033 × わかりやすさ − 0.015 × 古さ + 0.018 × 新しさ − 0.024 × 比喩性 + 1.965

付録の表 4 に短単位動詞に対して検討した推定代表義度上位・下位のものについて示す。表中「出現位置」に S がついているものはサ変動詞「為る」の用例のうち、サ変名詞が左に隣接するものについて、連結して呈示したものである。代表義度上位のものは、現代日本語で定着した外来語が多く見られた。代表義度下位のものは、「古さ」「比喩性」の評定値が高い用例が多くみられた。

付録の表 5 に多義の短単位動詞「掛かる」(語彙素番号：6016) の分類語彙表番号ごとの平均評定値（マクロ平均）を示す。用例数としては「.16 関係-時間」が 27 件と最も多く、代表義度も 2.114 と高い水準であった。一方、代表義度が最も高い「.31 活動-言語」は「電話がかかる」の意味だが、用例としては 1 事例のみであった。近年電話がかかってくることも少なくなっており、用例数が少なくなることが想定され、いずれ代表義度も低くなる可能性がある。

付録の表 6 に「掛かる」の代表義度上位・下位用例を示す。代表義度上位の用例は「.11 関係-類」および「.16 関係-時間」であった。代表義度下位の用例は「.3370 活動-生活-遊楽」（囲碁における用語)、「.1502 関係-作用-開始」（はじめる)、「.1513 関係-作用-固定・傾き・転倒など」（覆いかぶさる）などであった。興味深いことに、本来比喩性のない字義的な語義である PM29_00003 の例「それ（髪）が【かかっ】た肩先」(.1513 関係-作用-固定・傾き・転倒など）の代表義が低い。一方、抽象性の高い「.1110 関係-類-関係」「.1600 関係-時間-時間」「.3730 活動-経済-価格・費用」などが代表義度が高く、さらには本来比喩表現であるのにも関わらず比喩性が低い傾向が見られた。

6　おわりに

本研究では、コーパスの用例に対する印象評定データの構築を行った。BCCWJ の語義が付与された部分 (BCCWJ-WLSP) と読み時間が付与された部分 (BCCWJ-SPR2) に対して、短単位動詞・長単位・文節単位に、自然さ・わかりやすさ・古さ・新しさ・比喩性の評定値を収集した。さらに、短単位動詞については、得られた印象評定情報から代表義度を線形回帰により推定し、コーパス用例の典型用例抽出を試みた。

コーパスにおける出現度数と一般的な読み手の評定値とを対照することで、語の生産実体・受容実体を確認することができる。また、代表義度の推定や典型用例の抽出は、多義語における中心的な意味・基本的な意味の解明に寄与するほか、文法性・非文法性の判断にも寄与する。さらには、言語学習者に用例を提示するにあたり、典型用例を提示することが言語の流暢性の醸成に役立つと考える。

今後、BCCWJ-WLSP における語義との対照を行うことで、語義の転換が発生している箇所において、一般的な読み手が比喩性を感じるかなどを調査する。さらに、文節単位に付与された印象評定情報と読み時間とを対照することで、読み時間が変化する表現を印象評定情報の観点から解明したい。

6）山崎・柏野のデータにおける記号を、「●」:5,「○」:4,「△」:3,「×」:2, その他（「？」など）：1 とした。

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です。また、科研費18H05521, 18K18519、国立国語研究所第４期共同研究プロジェクト事前準備経費の支援を受けました。

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単語親密度データの構築 (2021 年版). In **言語処理学会第** *28* **回年次大会発表論文集**, 2022.

[3]Kikuo Maekawa, Makoto Yamazaki, Toshinobu Ogiso, Takehiko Maruyama, Hideki Ogura, Wakako Kashino, Hanae Koiso, Masaya Yamaguchi, Makiro Tanaka, and Yasuharu Den. Balanced corpus of contemporary written japanese. *Language Resources and Evaluation*, 48:345–371, 2014.

[4]天野 成昭 and 近藤 公久, editors. **単語親密度，*NTT* データベースシリーズ　日本語の語彙特性第 *1* 巻**. 三省堂, 1999.

[5]天野 成昭 and 近藤 公久, editors. **単語親密度増補，*NTT* データベースシリーズ　日本語の語彙特性第 *9* 巻**. 三省堂, 2008.

[6]藤田 早苗 and 小林 哲生. 単語親密度の再調査と過去のデータとの比較. In **言語処理学会第** *26* **回年次大会発表論文集**, pages 1037–1040, 2020.

[7]NTT コミュニケーション科学基礎研究所.

[8]佐久間 尚子, 伊集院 睦雄, 伏見 貴夫, 田中 正之, 天野 成昭, and 近藤 公久. ***NTT* データベースシリーズ 日本語の語彙特性 第 *8* 巻 単語心像性**. 三省堂, 2008.

[9]加藤 祥 and 浅原 正幸. 『現代日本語書き言葉均衡コーパス』出版書籍サンプルの NDC 別語彙分布語彙分布. In **言語資源活用ワークショップ** *2021*, pages 218–225, 2021.

[10]加藤 祥, 森山 奈々美, and 浅原 正幸. 『現代日本語書き言葉均衡コーパス』新聞記事情報を用いたジャンル別語彙分布. In **言語資源活用ワークショップ** *2021*, pages 160–170, 2021.

[11]加藤 祥 and 浅原 正幸. 多義語語義調査を目指した IPAL 形容詞例文への印象評定情報付与. In **言語処理学会第** *27* **回年次大会発表論文集**, pages 1120–1025, 2021.

[12]加藤 祥 and 浅原 正幸. Ipal 用言例文への印象評定情報付与と代表義・典型用例の抽出. **計量国語学**, 33(3):178–183, 2021.

[13]瀬戸 賢一. 多義記述の問題点とその解法–日本における正しい多義記述の出発点–. In **日本認知言語学会論文集**, pages 507–517, 2019.

[14]国広 哲弥. **意味論の方法**. 大修館書店, 1982.

[15]籾山 洋介. **認知意味論のしくみ**. 研究社出版, 2002.

[16]松本 曜. 多義性とカテゴリー構造. In 澤田 治美, editor, **ひつじ意味論講座**, volume 1, pages 23–43. 2010.

[17]瀬戸 賢一. メタファーと多義語の記述. In 楠見 孝, editor, **メタファー研究の最前線**, pages 31–61. ひつじ書房, 東京, 2007.

[18]山崎 誠 and 柏野 和佳子. 『分類語彙表』の多義語に対する代表義情報のアノテーション. In **言語処理学会第** *23* **回年次大会発表論文集**, pages 302–305, 2017.

[19]加藤 祥, 浅原 正幸, and 山崎 誠. 分類語彙表番号を付与した『現代日本語書き言葉均衡コーパス』の書籍・新聞・雑誌データ. **日本語の研究**, 15(2):134–141, 2019.

[20]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.

付録：

サンプル ID	出現位置	自	わ	古	新	比	例文	代表義度
代表義度上位								
PB35_00013	76720S	4.6	4.6	0.45	2	0.55	毎日【リラックスする】場所が、これだけ木の香りがすると最高です！	2.18805
PN1d_00005	2120S	4.65	4.65	0.8	2	0.5	まだキャンプ序盤ということで制球が定まらず、イチローが【スイングし】たのは２８球のうち７球だけ。	2.18625
PM41_00026	35240	4.6	4.65	0.9	1.95	0.4	どうやら齋藤アナ、オチャメ系女子アナと【し】ての自分をアピールするつもりらしい。	2.18565
PN5g_00006	620S	4.2	4.2	0.8	3.15	0.55	二十日投開票された長崎県雲仙市長選・市議選で、同市選管は、一部の不在者投票数を二重に【カウントし】たため、投票者数が三十一人多くなっていたとして、投票率を８４．６４%から８４．５７%に訂正した。	2.1855
PM25_00067	10350S	4.4	4.2	0.85	3.55	1	調査会社Ｊ．Ｄ．パワーアンドアソシエイツ（Ｊ．Ｄ．Ｐｏｗｅｒ ａｎｄ Ａｓｓｏｃｉａｔｅｓ）によると、携帯電話を使ってインターネットに【アクセスし】ている加入者比率は、２００１年の調査対象全体の２３パーセントで、前年の１２パーセントから２倍近い伸びを示している。	2.18355
代表義度下位								
PB39_00013	12710	1.9	1.65	2.9	0.75	3.2	数江が【しゃちほこばっ】てうなずく。	1.93545
PN5a_00003	15940	1.25	0.95	2.9	2.25	2.55	便乗批判受け新幹線【盛ら】ず	1.94715
PM42_00026	1150	1.7	1.35	3.25	0.65	1.75	こうして見てくると、「ノーベル文学賞」の生死など、現代史におけるほんのわき道の小エピソードとすら見えてくる次第ながら、やはり六〇年代末期から七〇年代半ばというのが、ノーベル文学賞に、ほとんど致命的な一撃をもたらしたという実感を【否み】がたいのだ。	1.9509
PB29_00003	27140	1.4	0.9	2.75	0.9	1.4	車を徐行させて運転手が左右を【ねめ回し】た時、志津がフロントガラスを見すえて叫んだ。	1.95285
PM41_00060	4310	1.4	1.3	2.2	1.4	2.45	白７６は堅すぎで、７８と【マゲ】て下辺を制限したい。	1.9581

表 4　推定代表義度に基づく典型用例の抽出（短単位動詞）

分類語彙表番号	自然さ	わかりやすさ	古さ	新しさ	比喩性	代表義度	用例件数
2:用	3.93	3.89	1.88	1.34	1.21	2.108	66
11:関係-類	3.92	3.85	2.15	1.37	1.22	2.102	10
1110:関係-類-関係	3.92	3.85	2.15	1.37	1.22	2.102	10
15:関係-作用	3.87	3.76	1.94	1.22	1.19	2.100	18
1502:関係-作用-開始	3.80	3.73	2.05	1.39	1.28	2.097	7
1513:関係-作用-固定・傾き・転倒など	3.92	3.78	1.87	1.11	1.13	2.102	11
16:関係-時間	4.00	3.99	1.77	1.36	1.19	2.114	27
1600:関係-時間-時間	4.00	3.99	1.77	1.36	1.19	2.114	27
31:活動-言語	4.20	4.30	2.10	1.80	1.50	2.122	1
3122:活動-言語-通信	4.20	4.30	2.10	1.80	1.50	2.122	1
33:活動-生活	2.35	2.25	2.75	2.00	2.20	2.009	1
3370:活動-生活-遊楽	2.35	2.25	2.75	2.00	2.20	2.009	1
37:活動-経済	4.02	4.05	1.67	1.43	1.13	2.120	8
3710:活動-経済-経済・収支	4.28	4.23	2.05	1.15	1.10	2.119	2
3730:活動-経済-価格・費用	3.93	3.99	1.54	1.52	1.14	2.121	6
51:自然-物質	3.85	4.00	1.55	0.95	1.55	2.100	1
5152:自然-物質-雲	3.85	4.00	1.55	0.95	1.55	2.100	1

表 5　短単位動詞「掛かる」語彙素番号:6016 の分類語彙表番号別評定値

サンプル ID	出現位置	自	わ	古	新	比	例文	代表義度
PB56_00007 WLSP:2.1110	41660	4.65	4.55	1.65	1.45	0.6	あいのり商法の成功はお互いに、ウイン・ウインの関係が構築できるかどうかに【かかっ】ているのだ。	2.1579
PB40_00003 WLSP:2.1600	15100	4.2	4.05	1.35	1.8	0.5	セミナーや講習会を受ける→時間が【かかる】→タイミングが合わない	2.1492
PB40_00003 WLSP:2.1600	4880	4.5	4.3	1.35	0.95	0.45	さがすのにも時間が【かかる】。	2.14695
PM41_00060 WLSP:2.3370	38420	2.35	2.25	2.75	2	2.2	第１譜、白２０と【カカっ】たのは戦いに自信のある表われ。	2.0094
PM25_00084 WLSP:2.1502	1310	2.4	2.1	1.85	1.9	1.5	下手したら回収に【かかっ】てるから。	2.03355
PB29_00003 WLSP:2.1513	6880	3	2.85	2.4	1.55	1.55	明け方にはこちらを向いていた顔が今は枕の向こうに落ち、解いた髪と、それが【かかっ】た肩先がこちらを向いている。	2.04975

表 6　短単位動詞「掛かる」語彙素番号:6016 の推定代表義度上位・下位用例