日本語母語話者にとっての英単語親密度の調査

藤田 早苗　　小林 哲生　　服部 正嗣　　納谷 太
NTT コミュニケーション科学基礎研究所
{sanae.fujita.zc, tessei.kobayashi.ga, hattori.takashi.cp, futoshi.naya.ke}@hco.ntt.co.jp

概要

我々は、主に学習支援への利用を目的とし、日本語母語話者にとっての英単語親密度を調査した。これまで、特に日本語母語話者にとっての英単語親密度データで大規模なものは存在しなかった。本調査では、世界最大規模の 12,398 語を対象とし、評定も、どの程度「見たり聞いたりするか」「知っている・意味がわかるか」「書いたり話したりできるか」の 3 観点とする詳しい調査をおこなった。評定者は日本語を母語とする TOEIC800 点以上の大学生・大学院生 42 人 (評定後のスクリーニング通過は 38 人) で、各自全ての語を評定した。さらに、先行研究との比較分析を行い、特に英語母語話者との違いを示した。

1　はじめに

単語親密度 (FAM; word familiarity) とは，語のなじみ深さを数値化したものである [1, 2]．単語親密度のデータは，さまざまな用途で使用されている [3, 4, 5]。特に、多くの語を含む単語親密度データがあると語彙数の簡便な推定が可能になるため [6]、学習支援への利用も期待できる [7]。本稿では、英語の学習支援への利用を目指して実施した日本語母語話者にとっての英単語親密度の調査について報告する。

英単語親密度のデータは幾つか存在するが、多くは英語母語話者による評定である [8, 9, 10](表 1, 上部)。*G&L*[10] と *GlasgowNorms*[8] はスコットランドの、*BristolNorms*[9] はイギリスの大学の学生・関係者が 7 段階のリッカート尺度で評定しており、約 1,500 語から 5,000 語程度のサイズである。

しかしながら、日本語母語話者にとっての英単語親密度が英語母語話者と同じような値になるとは限らない。日本の大学で調査したデータとしては、横川 [11] (以下、*Yokokawa*) によるものがある。関西の 10 大学の学生 822 人が 2,999 語の評定を行っているが、1 人が評定する語は 200 語のみである。大変参考になるデータだが、評定した大学生の英語力が分からない、語彙数推定の規準にするには調査語が少ない、同じ評定者によって全語評定されたわけではない、といった問題がある。

一方、単語親密度に限らなければ日本人英語学習者向けの単語リスト作成の試みは、多く存在する [12, 13] 。Ehara[12] は、ALC[1] が作成したレベル別語彙リスト SVL12000[2]（12,000 語）を対象に、意味を知っているかなどの評定を行っている。評定者は大学生 16 人 (15 人は東大生。日本語母語話者 14 人) である。規模が大きい上に全評定者が全語を評定している大変貴重なデータである。ただし、1-5 の評定値が振られているものの[3] 名義尺度のため、評定平均を算出して単語親密度として利用することはできない。

大学英語教育学会基本語リスト (新 JACET 8000) [13] は、主に教育者からなる委員会がコーパスや検定教科書を元に語を選定し、出現頻度等を基にランクを付与したものである。また、8,000 語からなる基本語リスト以外に、補正資料として、中学・高校コミュニケーション支援語彙リスト (3,000 語)、共通学術語彙リスト (2,200 語) も公開されている。大変丁寧に選定・編集されたデータである。

CEFR-J Wordlist ver.1.6 [4] は、6,868 語 (品詞別だと 7,989 項目) に対し、欧州共通言語参照枠 (CEFR) が定める 6 レベルのうち A1〜B2 の 4 レベルを付与した語彙リストである。各レベルは A1，A2 が基礎段階の言語使用者，B1，B2 が自立した言語使用者となっている。このリストは、中国・台湾・韓国の小

1） https://www.alc.co.jp/
2） https://www.alc.co.jp/vocgram/article/svl/
3） 各評定値と説明 : 1: never seen the word before, 2: probably seen the word before, 3: absolutely seen the word before but don't know its meaning/ tried to learn the word before but forgot its meaning, 4: probably know the word's meaning/ able to guess the word's meaning, 5: absolutely know the word's meaning
4） 東京外国語大学投野由紀夫研究室. http://cefr-j.org/index.html よりダウンロード

表 1　英語の親密度調査の比較

出典等	語数	評価者数/語	調査項目	補足
G&L [10]	1,944	36	見聞きしたり使う程度	スコットランド。名詞のみ
Bristol Norms[9]	1,526	20	[10] と同様	イギリス
Glasgow Norms[8]	4,682	平均 33	familiar さの程度 (知らない語だと “unfamiliar word” を選択)	スコットランド。一部は、語義ごとにも付与
横川 [11]	2,999	平均 54	見聞きする程度	200 語/人。 関西の 10 大学の学生で調査
本調査	12,399	38 (42)	3 軸（見聞きする程度など）	全語を同じ評価者が評定

中高の主力教科書をベースに構築されている。貴重なデータだが、4 レベルなので語彙数推定に利用するには情報量が足りず、かつ、必ずしも日本人のみ向けとは限らないという問題がある。

本稿では，日本語母語話者にとっての英単語親密度を調査する (2 章)。本調査では、評定者の英語力を統制するため、TOEIC800 点以上を条件とした。対象語は異なりで 12,398 語、各評定者は全語を評定する。42 人が全評定を実施したが、評定後に実施したスクリーニングを通過した 38 人分の評定値から英単語親密度を計算する。さらに、英語母語話者などの先行研究と特徴を比較・分析する (3 章)。

2　英単語親密度調査方法

評定者　評定者は TOEIC800 点以上の、日本語を母語とする大学生・大学院生である。TOEIC800 点以上としたのは、親密度に差をつけて評定をするためには、英語がある程度出来る必要があると考えたためである。

調査への応募者は 53 人だったが、全語の評定を終了したのは 42 人だった。さらに、調査後のスクリーニングを通過したのは 38 人 (男性 21 人、女性 17 人) だった。38 人の TOEIC の平均点は、896 点 (Listening 455, Reading 441) で、年齢は 18 歳から 25 歳、平均 21.5 歳だった。

調査対象語　学習支援への利用を目指し、英語の教科書コーパス（NTT で構築中。2018 年購入の小 5-高 3 の検定教科書全 72 冊から、物語と会話部分を書き起こしたもの）に出現する 4,955 語をまず対象とした。ただし、数字、記号、固有名詞、2 文字以下の語、日本語のローマ字表記 (“miso”, “kabuki” など) は除いた。

さらに先行研究との比較も視野に、新 JACET 8000[13]、*Yokokawa*、CEFR-J Wordlist ver1.6 に含まれる語は基本的に調査対象とした[5]。これらのリストに複合語が含まれる場合には、スペースで分割した語も調査対象とした[6]。また、Ehara[12] の公開データから、日本人 14 人による評定値の平均を便宜的に算出、平均値が 3.07 以上の 9,870 語は対象とした。

また、活用形による親密度の違いを見るために 100 語は活用形を追加した。さらに、スクリーニングのために 101 語を 2 回提示することとし、のべ 12,500 語 (異なり 12,398 語) を調査対象語とした。

対象語の提示順　調査語はランダムにリストにした。ただし、2 回提示する語は 1 回目提示から 3000 語以上離れて提示されるようにした。調査では、リストの昇順/降順に評定する人を半々とした。なお、事後スクリーニングを通過した 38 名のうち、20 名は昇順、18 名は降順からの提示だった。

調査手順　手順の概要は次の通りである．

Step0　インストラクション
評定開始前にオンライン説明会を実施。

Step1　練習フェーズ （10 語）
最初のログイン時に 1 度だけ実施。

Step2　本番フェーズ (12500 語) (付録の図 7 参照)
同じ評定が 15 回続くとアラートを表示。

調査後　事後スクリーニング
2 回評定させた語の重み付きカッパ係数 κ が 0.9 より低い評定者を除外。

ここで κ は、完全に評定が一致している場合を 1、評定が最も離れている場合を 0、評定が完全一致でなくともより近い値なら重みを重くして一致度を計算する。

謝金は全ての評定実施とアンケート提出で 4 万円とした。評定サイトは内製し、スマホでも PC でも評定結果は自動保存するようにした。調査期間は、2021.6.28 - 2021.8.16 である。評定のペースは評定者自身に任せた。

評定軸　同じ「親密度（familiarity）」と名付けられていても、評定者への尋ね方は先行研究によって

5）’s, ’m, ’re, 擬似語等は除いた

6）例えば、“air force” の場合、“air” と “force” も対象語に追加する

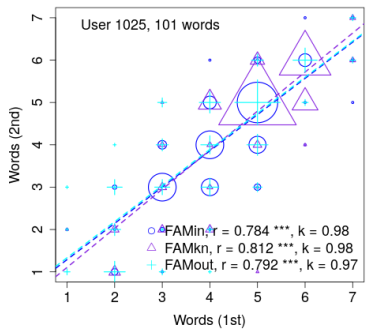

図 1　一貫性評定例 1: 利用

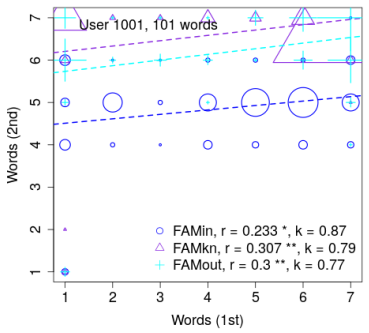

図 2　一貫性評定例 2: 除外

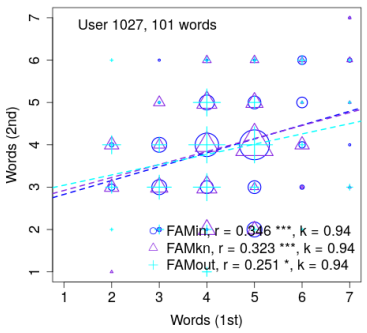

図 3　一貫性評定例 3: 利用

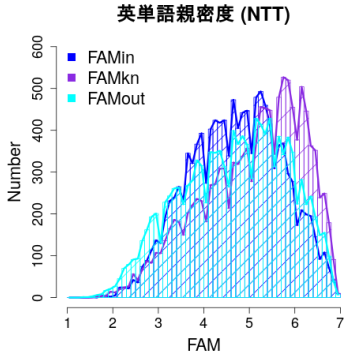

図 4　ヒストグラム:全調査

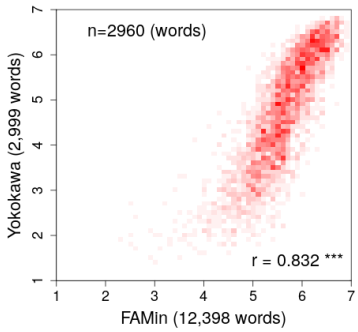

図 5　FAM_{in} と $Yokokawa$

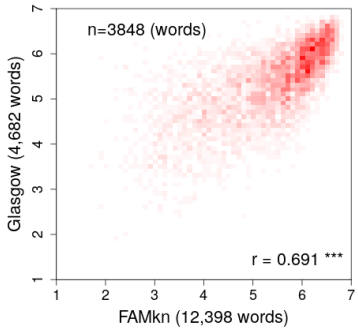

図 6　FAM_{kn} と $GlasgowNorms$

異なる (表 1)。例えば、$G\&L$ は見聞きしたり使う程度を尋ねているのに対し、$GlasgowNorms$ は、なじみ (familiar) の程度、$Yokokawa$ は見聞きする程度を尋ねている。

尋ね方によって、評定結果は異なってくると考えられる。そこで本調査では、3 通りの尋ね方で評定を実施した。全ての評定者は、全ての語について、「どの程度見たり聞いたりするか (FAM_{in})」「どの程度知っている・意味が分かるか (FAM_{kn})」「どの程度書いたり話したりできるか (FAM_{out})」という 3 つの観点で 7 段階評定を実施する。

なお、FAM_{in} は $Yokokawa$ と同じ観点である。$G\&L$ や $GlasgowNorms$ と完全一致する観点はないが、あえて言うなら前者は FAM_{in},FAM_{out} の混合、後者は FAM_{kn} に近いだろう。

3　結果と分析

評定者ごとの結果とスクリーニング　図 1 - 3 にスクリーニングに用いた重複語の評定結果の例を示す。図 1 - 3 には、重複語の重み付きカッパ係数 κ だけでなく pearson の相関係数 r も記載した。図 1, 3 は評定結果を利用、図 2 は除外した例である。図 1 の評定者は κ も非常に高く、r も 0.784〜0.812 と強い相関がある。一方、図 2 の評定者は、κ も規準以下で相関も弱い。図 3 の評定者は、r だけをみると相関は弱いが、κ は基準値を超えている。r が低いのは、評定値が真ん中あたりに集まっている事が原因の一つと考えられる。

スクリーニングの方法と規準は色々考えられる。日本語の令和版単語親密度のクラウド調査 [2] では、$r \geq 0.5$ を閾値としている。ただし、令和版単語親密度調査では、重複語は全体の 5%であり本調査よりかなり多い。のべ 12,500 語のうち 5%を重複させるなら 595 語を重複させる必要があるが、本調査では 101 語のみ重複させた。重複語が比較的少ない場合、図 3 のように、一致度は高いが r は高くないという場合があり得る。そのため本調査では、より一致度を重視して κ を閾値として採用した。

本調査で事後スクリーニングを通過した人は、42 人中 38 人 (90.5%) だった。比較のために、$r \geq 0.5$ を閾値として計算してみると通過率は 69.0%となる。これは、日本語の令和版単語親密度 [2] のクラウド調査時の通過率 (35.6%) よりはるかに高い。なお、実験室に評定者を集めて調査した平成版の単語親密度 [1] では通過率が 80% だった。本調査はオンライン調査で、かつ、12,500 語と評定語が多いにも関わ

表 2　各単語親密度データとの相関

	一致語数	FAM_{in}	FAM_{kn}	FAM_{out}
BristolNorms	1,325	0.628	0.602	0.622
GlasgowNorms	3,848	0.708	0.691	0.708
Yokokawa	2,960	**0**.832	0.787	0.818
FAM_{kn}	12,398	0.977	-	-
FAM_{out}	12,398	0.988	0.989	-

いずれも $p < 0.05$

らず、高い通過率となった。この理由として、オンライン説明会を実施するなどお互い顔の見える状態で依頼をしたことや、一般的なクラウド調査より高い謝金設定としたこと等が考えられる。以降の分析では、この 38 人の評定の平均を利用する。

評定軸同士の比較　本調査結果のヒストグラムを図 4 に示す．ヒストグラムの形を比べると、FAM_{kn} はピークが 5.8 程度と最も高い。一方、FAM_{out} のピークは 5 程度で、形もよりなだらかである。FAM_{in}〜FAM_{out} で差が大きい語は少ないが、例えば FAM_{in} と FAM_{kn} で 1 以上差があったのは “policewoman”,“mountaintop” など 57 語だった。いずれも FAM_{kn} の方が FAM_{in} よりも高く、あまり見聞きしない語でも分かる場合はあるが、よく見聞きする語が分からない場合は少ないことが読み取れる。

先行研究との比較　表 2 に各先行研究との間の相関係数を示す。各先行研究が同じ語に対する調査ではないので必ずしも公平な比較ではないが、日本人で調査した *Yokokawa* とは強い相関があり、特に、同じ「見聞きする程度」を調査した FAM_{in} と最も強い相関が見られた。一方、共通の語の散布図 (図 5) からは、*Yokokawa* の方が値が低い傾向が見て取れる。これは *Yokokawa* の調査では大学生の英語力を統制していないのに対し、本調査では TOEIC800 点以上と統制したためだと考えられる。

英語母語話者での調査と比較しても中程度〜強い相関が見られる。*GlasgowNorms* と FAM_{kn} で共通する語の散布図を図 6 に、両者で差が大きい語を表 3 に示す．図 6 からは、*GlasgowNorms* の方が若干高めな語は多いが、*Yokokawa* より対角線上付近の語が多い、つまり、数値が近い語が多い傾向が見て取れる。さらに表 3 から、FAM_{kn} の方が高い語には、“carp” や “mermaid” など日本でもカタカナでよく使われる語が多く、日本語の影響が見て取れた。一方、FAM_{kn} の方が低い語には、“duvet” や “plaster” のような身の回りのものや、“wee” のような子どもであればよく使うであろう語が多く含まれていた。こうした違いは母語か第 2 言語かに拠るところが大きいと考えられる。

表 3　*GlasgowNorms* と FAM_{kn} で差分が大きい語

FAM_{kn} の方が高い

No.	語	Glasgow	FAM_{kn}	差分
1	carp	2.515	4.92	-2.405
2	taboo	3.647	5.92	-2.273
3	mermaid	3.852	5.95	-2.098
4	wagon	3.482	5.58	-2.098
5	seldom	4.031	6.05	-2.019
6	pioneer	4	5.97	-1.97
7	anthropology	3.143	4.97	-1.827
8	phenomenon	4.147	5.97	-1.823
9	mall	3.815	5.58	-1.765
10	paradox	4.029	5.79	-1.761

FAM_{kn} の方が低い

No.	語	Glasgow	FAM_{kn}	差分
1	duvet	5.964	2	3.964
2	plaster	6.258	2.68	3.578
3	wee	5.758	2.26	3.498
4	parsley	5.774	2.47	3.304
5	lad	5.758	2.47	3.288
6	granny	6.156	2.87	3.286
7	horrid	5.257	2.26	2.997
8	hen	5.909	2.92	2.989
9	celery	5.645	2.68	2.965
10	bloke	5.647	2.82	2.827

4　まとめと今後の課題

本稿では，オンライン調査による信頼度の高い親密度評定方法を提案、日本語母語話者にとっての英単語親密度を調査・分析した。調査した語は異なりで 12,398 語と、英単語親密度調査としては過去最大規模である。評定は 42 人完了したが、最終的に事後スクリーニングを通過した 38 人分の評定値の平均を用いた。評定者は TOEIC800 点以上とした。さらに、構築した英単語親密度データを先行研究と比較し、特に英語母語話者のデータとの比較により、日本語の影響や母語かどうかの違いによる影響を示した。

今後は、構築した英単語親密度データを用いた英語の語彙数推定や、語彙数と読解力との関係調査にも取り組みたい。

謝辞

この成果の一部は、国立研究開発法人新エネルギー・産業技術総合開発機構 (NEDO) の委託業務の結果得られたものです。

参考文献

[1]天野 成昭 and 近藤 公久. **日本語の語彙特性**. 三省堂, 東京, 1999.

[2]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** *(NLP-2020)*, number F4-4, 2020.

[3]若松 千裕, 石合 純夫, 林 圭輔, and 相原 伸子. 語義聾症例における文字言語の理解過程 -通常見かけない文字表記語による検討-. **高次脳機能研究 *(*旧 失語症研究*)***, 36(1):9–19, 2016.

[4]水野 りか and 松井 孝雄. 文字の漢字表記語の意味処理に対する構成漢字の影響と処 理順序. **心理学研究**, 90(2):201–206, 2019.

[5]高橋 三郎. 学齢期の吃音児における語の長さが吃音頻度に及ぼす影響. **音声言語医学**, 59(2):188–193, 2018.

[6]藤田 早苗, 菅原 真悟, 小林 哲生, 新井 庭子, 山田 武士, and 新井 紀子. 小学生から高校生に対する語彙数調査と単語親密度との関係分析. In **言語処理学会第 *26* 回年次大会発表論文集** *(NLP-2020)*, number E1-3, 2020.

[7]藤田 早苗, 服部 正嗣, 小林 哲生, and 納谷 太. 日本人英語初学者の語彙数推定方法の検討. In **第 *34* 回人工知能学会全国大会** *(JSAI-2020)*, 2020.

[8]Graham G. Scott, Anne Keitel, Marc Becirspahic, Bo Yao, and Sara C. Sereno. The glasgow norms: Ratings of 5,500 words on nine scales. *Behavior Research Methods*, 51:1258–1270, 2019.

[9]Hans Stadthagen-Gonzalez and Colin J. Davis. The bristol norms for age of acquisition, imageability, and familiarity. *Behavior Research Methods*, 38:598–605, 2006.

[10]K. J. Gilhooly and R. H. Logie. Age-of-acquisition, imagery, concreteness, familiarity, and ambiguity measures for 1,944 words. *Behavior Research Methods & Instrumentation*, 12:395–427, 1980.

[11]横川 博一. **日本人英語学習者の英単語親密度 文字編**. くろしお出版, 2006.

[12]Yo Ehara, Nobuyuki Shimizu, Takashi Ninomiya, and Hiroshi Nakagawa. Personalized Reading Support for Second-Language Web Documents by Collective Intelligence. In *the 2010 International Conference on Intelligent User Interfaces (IUI-2010)*, pages 51–60, 2010.

[13]大学英語教育学会基本語改訂特別委員会, editor. **大学英語教育学会基本語リスト 新 *JACET8000***. 桐原書店, 2016.

付録 (Appendix)

図 7　調査画面