単語親密度と単語頻度の関係に関する一考察

寺田 博視　　田中 久美子
東京大学大学院 情報理工学系研究科
terada@cl.ci.i.u-tokyo.ac.jp, kumiko@i.u-tokyo.ac.jp

概要

本稿では，さまざまなコーパスから得た頻度と，認知実験を通して得られた親密度の関係を分析した結果を報告する．親密度と頻度の相関は，英語では 0.57 から 0.74，日本語では 0.45 から 0.72 であり，さらにつぎの 3 つの知見を得た．第一に，単語の頻度が大きいことは単語が親密であることの必要条件であるが，十分条件ではない．第二に，コーパスが大きいと相関も高くなる．第三に，話し言葉コーパスは書き言葉コーパスよりも相関が高い傾向にある．これらの分析結果は，日常的な内容の大規模なコーパスから得た単語頻度が，単語親密度の一尺度となる可能性を示唆する．

1　はじめに

単語親密度はある単語がどの程度なじみ深く感じられるかを表す指標である．例えば,「出会い」と「邂逅」はほぼ同じように用いられるが，明らかに前者のほうが易しくなじみ深い．単語親密度は言語認知過程を解明する手がかりとなる可能性を秘めていることから，これまでに言語心理学の分野で研究されてきた．その一貫として，様々な親密度リストが心理実験を通して構築されてきた (1) (2) (9)．とはいえ，親密度がどのような心理上の要因によって決まるのかは解明されたとは言い難い．

頻度が言語認知に重要な役割を果たしていることは，過去の研究が指摘しており (8) (3) (6)，親密度の要因であることも示唆されている (4)．この点，大量の言語データが入手可能となった昨今では，コーパスを用いて頻度を計測し，大規模な検証を行うことができる．そこで本稿では，心理実験で得られた親密度とさまざまなコーパスから取得した頻度の関係を調査し，その結果を報告する．

単語の難易度は，文書の難易度の関わる．このため，本研究の成果は，教育の観点からの言語工学に応用することができる点で，工学的にも興味深い．本稿では，まず基本的な相関に関する知見を示した後，コーパスの大きさや種類がどのように影響するのかを調べる．

2　データ

2.1　単語親密度リスト

英語の親密度リストは歴史が長く，1950 年以前から研究が行われている (7)．その後，複数の親密度リストがまとめられ，MRC 親密度リストとして公開された (2)．MRC リスト[1] には 26 種類の属性値を付与した 150,837 語が収録されている．とはいえ，すべての単語に全属性が付与されているわけではなく，親密度が付与されている単語は 4,894 語のみである (以下，MRC リストとする)．日本語では，天野らが大規模実験を通して 68,550 語の単語親密度を得ている (1)(以下，Amano リストとする)．本研究ではこの２つのデータを親密度データとして利用する．

リストに収録されている単語は，1.0（なじみがない）から 7.0（なじみがある）の範囲の親密度が実数値で表されている[2]．Amano リストは内容語のみで構成されているのに対し，MRC は機能語を含んでいる．

2.2　コーパス

頻度を計測するためにコーパスが必要となる．本研究で用いたコーパスを表 1 に示した．上段には英語のコーパスを，下段には日本語のコーパスを載せた．コーパスの種別とすると，新聞，Wikipedia，ウェブデータがあり，内容的には書き言葉，話し言葉，またそれらの混合データを用意した．

新聞は，3 年分の WSJ コーパス (英語) と 5 年分の毎日新聞コーパス（日本語）を用いた．Wikipedia は，英語と日本語の平文を抽出し，それぞれ 1,912,595 ページ（4.74GB），372,890 ページ (956MB) を取得した．Web-E や Web-J はウェブデータであり，2006

[1] http://www.psy.uwa.edu.au/mrcdatabase/uwa_mrc.htm
[2] MRC は本来，100 から 700 で表されているが，Amano リストに合わせるため本来の値を 100 で割った値を用いた．

表 1: 単語頻度を計測するために用いたコーパス/単語親密度と単語頻度の相関

ラベル	単語数	異なり単語数	大きさ	話し言葉/書き言葉	種類	Pearson	Spearman
WSJ	42287431	127353	249 MB	書き言葉	新聞	0.6520	0.7082
Wikipedia-E	711143194	168533	4.7 GB	書き言葉	百科事典	0.6677	0.6801
Web-E	88267343947	204724587	1.9 TB	書き言葉	全般	0.7185	0.7359
BNC	97098970	364262	270 MB	混合	全般	0.7438	0.7776
MICASE	1279792	—	—	話し言葉	学術講演	0.5744	0.7127
Mainichi	80709011	198767	473 MB	書き言葉	新聞	0.6327	0.5330
Wikipedia-J	130418600	619636	956 MB	書き言葉	百科事典	0.5737	0.4452
Web-J	7183558565	5474644	69 GB	書き言葉	全般	0.7193	0.4920
Aozora	25975560	139961	172 MB	書き言葉	文学	0.4484	0.3451
CSJ	7498763	47767	40 MB	話し言葉	学術講義	0.4931	0.5097

年秋にインターネットから収集されたデータ[3]を用いた．英語と日本語はそれぞれ，265,823,502 ページ (1.9TB)，12,751,271 ページ (69GB) に及ぶ．書き言葉と話し言葉が含まれた英語のコーパスとして，British National Corpus(BNC) を用い，日本語のコーパスとして，青空文庫 (Aozora)[4]を用いた．話し言葉の英語のデータとして，Michigan Corpus of Academic Spoken English(MICASE)[5]を用い，日本語のコーパスとして，日本語話し言葉コーパス (CSJ)(5) を用いた．

3 基本的な相関

頻度の値域を親密度にあわせるために，頻度は対数を用いる．頻度の対数は，頻度確率で単語の出現確率を近似した際に情報量との関係が深いため，頻度の対数と親密度の関係を調べることは，単語の情報量と親密度の関係を調査することであると解釈することができる．

WSJ と Web-E における親密度と頻度 (対数) の散布図を図 1 に示した．横軸は親密度，縦軸は頻度 (対数) である．一般に，点は左下から右上に位置するが (右図)，形状は右下に鈍角のある三角形となることがある (左図)．これは，低頻度にもかかわらず，親密度の高い単語が多く含まれていることを示しており，特に，コーパスが小さいときに，三角の形状が際立つ．そのような低頻度・高親密度の単語の例を，以下に挙げる．

spank (親密度=5.36 / 頻度=19),
pimple (5.57 / 20), dime (5.86 / 39),
easygoing (5.25 / 41), quart (5.68 / 50)

一方，左上の領域には，単語はほとんど皆無である．つまり，高頻度であるが低親密な単語はほぼないということである．以上から，頻度の大きい単語は親密度が高いが，その逆はいえず，頻度が高いことが親密であることの必要条件となっている．

どの程度相関しているかを検証するために，ピアソンのモーメント相関係数とスピアマンの順位相関係数を算出した．ピアソンの相関係数は欠損値を考慮せずに計算されるが，一方，スピアマンの相関係数は欠損値を考慮して計算されるという差がある．結果を表 1 の右端の 3 列に示す．

ピアソンの相関係数の高いコーパスはスピアマンの相関係数も高い傾向にある．また，相関はコーパスを大きくすると高くなる．その理由として，コーパスが大きければ，得られた頻度の信頼性も向上するからと考えられる．実際，コーパスが大きいウェブデータのピアソン相関係数は他よりも高い．また，コーパスを大きくすると，散布図の形状は三角形から帯状へと変化する．たとえば，図 1 の右図は，大きさは 270MB であるのに対し，左図は 249MB である．それゆえ，コーパスの大きさは相関に影響を与える要因のひとつである可能性が高い．

とはいえ，相関に影響を与える要因はコーパスの大きさだけとはいえない．例えば，BNC は Web-E よりも規模が小さいにもかかわらず，相関は BNC のほうが高い．日本語においても，Mainichi と Wikipedia で同様のことが言える．その原因の可能性として，コーパスの種類の問題が挙げられるだろう．高相関のコーパス (BNC, ウェブデータ, 新聞) と低相関の

[3]東京大学の田浦研究室が収集したデータを用いた
[4]http://www.aozora.gr.jp/
[5]http://quod.lib.umich.edu/m/micase/

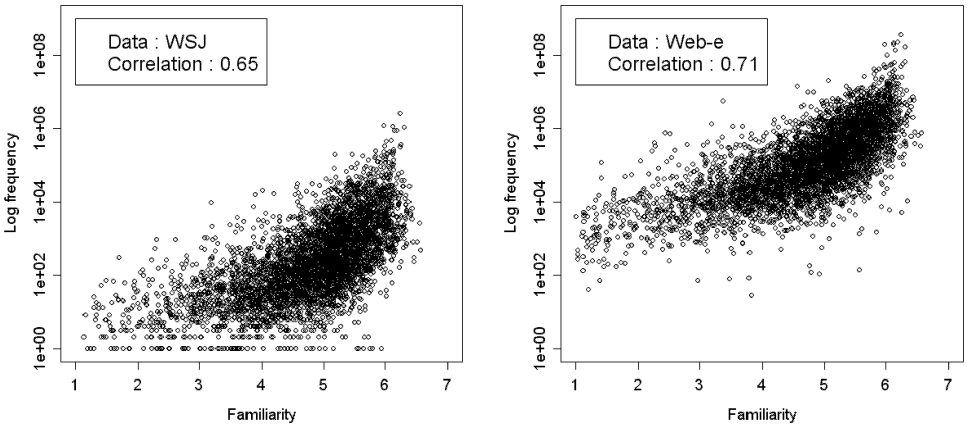

図 1: WSJ(左) と Web-E(右) における単語頻度と単語親密度の散布図

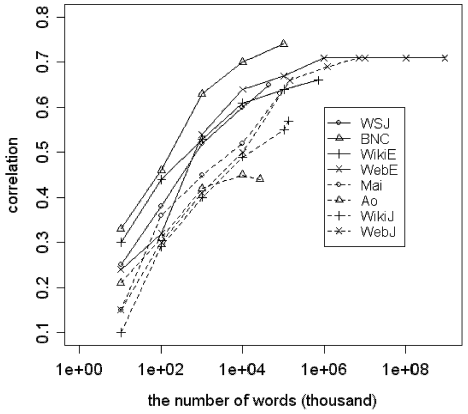

図 2: コーパスの大きさ (単位：千単語) と相関係数

コーパス (青空文庫,Wikipedia) の内容を比較すると，より日常的な内容を含んでいるほうが相関が高い傾向にある．たとえば，BNC は高い相関を示しており，内容には話し言葉に近いテキストを多く含んでいる．一方で，WSJ は低相関であり，内容は書き言葉のみで形成されている．

以上まとめると，相関に影響を与える要因として次の 2 つが挙げられる．

- コーパスの大きさ：コーパスが大きいほど親密度との相関は高い
- コーパスの種類：話し言葉は書き言葉よりも相関が高い

次章以降では，これらをさらに検証する．

4 コーパスの大きさによる影響

コーパスの大きさと相関係数の関係を調査するために，同種のコーパスの大きさを変化させたときの相関を調べる．それぞれのコーパスから単語数を指数的に増やしたコーパス (10^1, 10^2, 10^3 ...) を標本抽出する．標本数は一万語からそれぞれのコーパスの大きさいっぱいまで増加させる．データの大きさと相関係数の関係を 図 2 に示した．横軸は標本数（対数），縦軸は相関係数を表している．図上の折線は，実線が英語コーパスで破線が日本語コーパスを指している．すべてのコーパスにおいて，約 10 億語あたりまで対数線形的に相関は大きさとともに増加していることがわかる．

5 コーパスの種類による影響

コーパスの大きさを固定してみると，相関には差がある．その要因を調べるために，コーパスの種類による影響を調査した．

まずは，低相関の要因となる単語として，高親密・低頻度である単語を抽出すると，pencil や noisy など，日常的に用いられている語や話し言葉を多く含んでいることがわかった．一方，低親密で比較的高頻度の単語には，hypothesis や essence など，書き言葉として用いられている単語が多く存在した．

そこで，話し言葉と書き言葉コーパスの相関係数を比較してみた．コーパスの大きさは相関に影響を与えるので，コーパスの大きさを揃え，また，親密度リストの単語数も相関に影響する可能性があるため，親密度の高い上位 N 語だけを用い，コーパスとの相関を調査した．

話し言葉と書き言葉コーパスの違いが影響している可能性があるため，BNC は話し言葉部分と書き言葉部分に分離した．結果，英語では，話し言葉が 2 コーパスと書き言葉が 5 コーパス，日本語では，話し言葉が 1 コーパス，書き言葉が 4 コーパスを検証することになる．

英語と日本語の結果をそれぞれ図 3，図 4 に示した．横軸は N で，縦軸は相関を表している．破線は

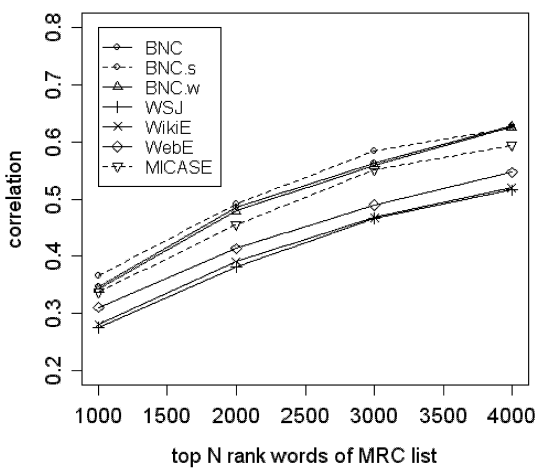

図 3: MRC リスト上位 N の単語と相関係数

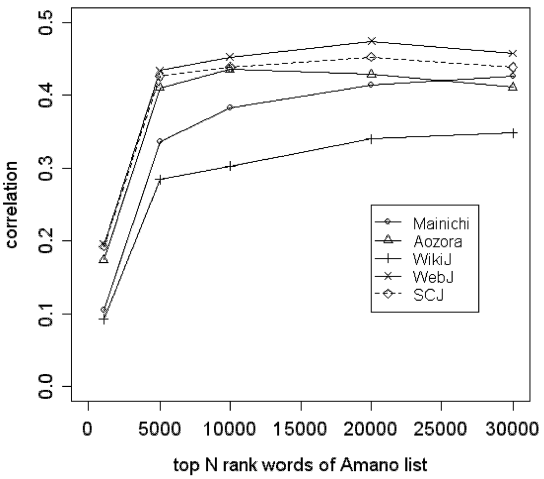

図 4: Amano リスト上位 N の単語と相関係数

は話し言葉，実線は書き言葉に対応する．破線は上方に位置し，書き言葉よりも話し言葉のほうが親密度との相関が高くなる傾向を示した．とはいえ，BNC の書き言葉は MRC よりも上に位置しており，必ずしも話し言葉のコーパスの相関が高いわけではない．その原因として，MICASE は話し言葉とはいえ，学術講演を収録しているのに対し，BNC の書き言葉には，日常的な内容の文書が含まれるためと考えられる．とすると，話し言葉，書き言葉に分離するよりも，日常的な内容かどうかなどで調査を進める可能性もあるが，現段階ではコーパス内にそのようなタグはないため，これは今後の課題となっている．

6 結論

本稿では，英語と日本語におけるさまざまなコーパスを用いて，単語親密度と頻度の関係を分析した結果を報告した．

頻度（対数）と親密度の相関は，英語では 0.57 から 0.74 で，日本では 0.45 から 0.72 であった．高頻度の単語は常に高親密であるが，高親密の単語は必ずしも高頻度ではないことが知見として得られた．

コーパスによって相関に差が生じる理由を説明するために，2 つの要因を調査した．第一の要因は，コーパスの大きさであり，大きいコーパスは小さいものに比べて強く相関した．コーパスの大きさを変化させることによって，相関は 10 億万語あたりまで対数線形的に増加することもわかった．第二の要因は，コーパスの種類の影響である．特に，話し言葉は書き言葉よりも強く相関する傾向がみられた．

以上から，日常的な内容の大規模なコーパスがあれば，そこから得た単語頻度は，単語親密度の擬似的な尺度となりうることを示唆している．今後は，以上の知見を文書難易度判定などに応用することを考えていきたい．

参考文献

[1] S. Amano and T. Kondo. On the ntt psycholinguistic databases ‘lexical properties of japanese’. *Journal of the Phonetic Society of Japan*, Vol. 4, No. 2, pp. 44–50, 2000.

[2] M. Cortheart. The MRC psycholinguistic database. *Quaterly journal of experimental psychology*, Vol. 33, No. A, pp. 497–505, 1981.

[3] E. Dupoux and J. Mehler. Monitoring the lexicon with normal and compressed speech: Frequency effects and the prelexical code. *Journal of Memory & Language*, Vol. 29, pp. 316–335, 1990.

[4] M.A. Gernsbacher. Resolving 20 years of incosistent interactions between lexical familiary and orthography, concretness, and polysemy. *Journal of Experimental Psychology: General*, Vol. 113, pp. 256–281, 1984.

[5] S. Furui K. Maekawa, H. Koiso and H. Isahara. Spontaneous speech corpus of japanese. 2000.

[6] Andrew Meade. Mark Pagel, Quentin D Atkinson. Frequency of word-use predicts rates of lexical evolution throughout indo-european history. *Nature*, Vol. 449, No. 7163, pp. 717–720, 2007.

[7] H. C. Nusbaum, D.B. Pisoni, and C.K. Davis. Sizing up the hoosier mental lexicon: Measuring the familiarity of 20,000 words. *Research on speech perception, progress report*, Vol. 10, pp. 357–376, 1984. Indiana University.

[8] J. Segui, J. Mehler, U. Frauenfelder, and J. Morton. The word frequency effect and lexical access, 1982.

[9] M. Wilson. Mrc psycholinguistic database: Machine-usable dictionary, version 2.00. *Behavior research methods, instructions, and computers*, Vol. 20, pp. 6–10, 1988.