Webコンテンツの文章表現に関する一検討

―単語親密度を用いた文章の親密度評価―

岡田　仁之† 島田　諭† 福原　知宏†† 佐藤　哲司†

† 筑波大学大学院 図書館情報メディア研究科　〒305–8550 茨城県つくば市春日 1–2
†† 東京大学人工物工学研究センター　〒277–8568 千葉県柏市柏の葉 5–1–5
E-mail: †{hito,sat,satoh}@slis.tsukuba.ac.jp, ††fukuhara@race.u-tokyo.ac.jp

あらまし　今日，Web 上では膨大な量のテキストが生み出されている．その中で，テキストを介したコミュニケーションにおいて齟齬や行き違いが増えており，書き手の意図が伝わる文章表現が求められている．文章表現を読み手に合わせて変換するには，読み手がテキストの印象を決定する要因を明らかにする必要がある．本稿では，テキストの印象を決定する重要な要因としてテキストの難易度を取り上げ，単語の難易度からテキスト全体の難易度を推定する手法を提案する．単語の難易度として，被験者実験に基づいて日本語語彙の認知度を網羅した単語親密度を用い，「朝日新聞」記事データ及び質問回答サイト「Yahoo! 知恵袋」データを用いて検討を行った．

キーワード　テキストベースコミュニケーション，難易度，日本語の語彙特性

A Study on Document Representation of Web Contents

—An Analysis of Readability of Documents using the Word Familiarity Metric—

Hitoyuki OKADA†, Satoshi SHIMADA†, Tomohiro FUKUHARA††, and Tetsuji SATOH†

† Graduate School of Library Information and Media Studies, University of Tsukuba
1-2, Kasuga, Tsukuba, Ibaraki 305–8550 Japan
†† Research into Artifacts, Center for Engineering, The University of Tokyo
Kashiwanoha 5–1–5, Kashiwa-shi, 277–8568 Japan
E-mail: †{hito,sat,satoh}@slis.tsukuba.ac.jp, ††fukuhara@race.u-tokyo.ac.jp

Abstract　Word Familiarity, Intelligibility Analysis, Text Mining, CGM

Key words　Today, text-based communications are still main on the WWW. However, as the diversity of the user extends, failures of the communication between users has increased, too. One of the causes is that the impression of the text differs among the author and readers. As a result, the difficulty of the text might needlessly become high. In this paper, we report on the result of examining the method for presuming the difficulty of text. We analyzed articles on "Asahi Shimbun" and question answer site "Yahoo! Chiebukuro (bag of knowledge)" by using word familiarity that was the actual measurement value that had been obtained by the testee experiment.

1. はじめに

今日，インターネットを介したコミュニケーションが浸透している．中でも，電子メール・電子掲示板・チャットなど，これらのほとんどはテキストを用いた意思疎通であり，従来に比べてテキストを記述する機会は格段に増えた．

テキストによる意思疎通と，対面での意思疎通との大きな違いとして，非言語情報の欠落が挙げられる．対面の意思疎通の場合，我々は意識的・無意識的に関わらず，相手の身なり，表情，手ぶり，声など，多くの非言語情報を基に，相手やその場の印象を受け取り，会話に反映させる．テキストによる意思疎通では，このほとんどを受け取れないため，対面と比べて，時に文脈を取り違えたり相手の機嫌を損ねる発言をしてしまったりする事が起こりやすい．メラビアンの法則[1]でも,「好感の合計=言葉による好感 7 %＋声による好感 38 %＋表情による好感 55 %」とされており，印象伝達において言語情報が占める割合の低さが読み取れる．

電子掲示板などでは，そもそも誰と意思疎通するのかわからない場合も多い．相手は不特定多数である．メーリングリストでは，情報の公開範囲は限定されているが，メッセージは参加

している複数人が受け取る．置かれる状況の異なる他者によって，メッセージが異なって認識され，誤解が広がる事も度々起こりうる．

以上のことから，インターネット上でのテキストを用いたコミュニケーションを円滑に行うには，送信するメッセージを慎重に吟味する必要がある．対面では相手の状況を推し量る情報は数多いが，テキストでは非常に限られた情報からそれを行わなければならない．本稿ではこの問題に対する基礎検討として，新聞記事ならびに Yahoo!知恵袋のデータを用いて，文書中に出現する単語の親密度に関する分析を行った．

本稿の構成は次の通りである．2. ではテキストを介するコミュニケーションにおける文書表現に対して考察する．3. で単語親密度を用いた分析について説明し，新聞記事を用いて予備的な分析を行った後，4. で Yahoo!知恵袋のデータを用いた分析と考察を行い，5. でまとめを述べる．

2. テキストを介したコミュニケーションにおける文書表現

例えば，遅刻する，という内容を表す文面でも，それを伝える相手によって様々な言い方が存在する．

（1） ごめん，ちょっと遅れる
（2） 申し訳ありません，今しばらくお待ち下さい
（3） 申し訳ありません，遅れます
（4） 体調が悪いので様子を見ております

どれも遅れる旨と，謝意が表わされているが，受ける印象はかなり異なる．(1) は相手が親しい友人の場合，(2) は目上が相手の場合や公の場である場合だと推測できる．また，この場合相手がどれ程迷惑に思っているか，機嫌を損ねるかなど，相手の状況を推し量って使う言葉をコントロールする必要がある．(4) は他と変わった言い回しで誤魔化している感があるが，目論見が成功するかは相手の状況次第である．この様に時と場所，相手に応じて文章を作成する作業は負担が大きい．

この様な負担を支援するために，文脈に応じた文章を推薦するシステムが考えられる．言葉選びの支援として，類語辞書を用いた選択候補の提示システムが既に開発されている．しかし，そもそも言葉選びの負担は，言葉の意味的な違いによるものではなく，言葉に対する話し手と受け手の概念・印象の違いによるものである．類語辞書を参照して話し手が持つ概念を正確に表現する語を選択したとしても，受け手がその語の概念を誤解していては齟齬が生じる．「今すぐには無理だからなし崩しにやっていこう」と言った場合，意味を正反対に捉える者がいるであろう．辞書は多くの人の言葉の概念の共通項を抜き出したものであり，最低限の意味は伝わる可能性が高いが，日常のコミュニケーションでは不十分である．更に強い概念・印象のマッチングを図るための言葉選びが必要とされている．

新聞等のマスメディアでは，言葉選びのガイドラインが用字用語集としてまとめられている．用字用語集は不特定多数を相手とする状況に限られている．世相にあわせて言葉は移り変わるため，頻繁に改定される．Web から自動的に用例を収集し，文脈に応じた候補を随時更新するシステムが出来れば，用字用語集に代わって言葉選びの作業を支援する事ができる．本研究では，この様なシステムの実現を目標として，Web の文書表現の基礎検討を行っている．

表 1　辞書に含まれる単語数

品詞	親密度辞書	IPADIC	共通
名詞 (サ変)	59,841	12,456	7,455
動詞	5,080	130,750	3,944
副詞	1,572	3,032	1,091
形容動詞	985	3,328	364
形容詞	792	27,210	585

本稿では，まず質問回答サイトにおける質問と回答，および新聞記事の各記事の文章表現の違いを調査した．新聞記事は，複数の人間によって書かれた文書表現であるが，強く統制されており，記事の種類によって難易度や文体は異なるとの仮説を置いた．今回は「単語に対する『なじみ』の程度を表す主観的評価値」[6], [8] である単語親密度を用いて，各記事を分析した．

3. 文章表現の分析

3.1 単語親密度

天野らは,「単語に対する『なじみ』の程度を表す主観的評価値」[6], [8] を として「単語親密度」を提案している．単語親密度とは「各単語を見聞きする経験の多さ，熟知度，意味理解のしやすさなどの総合的な指標」である．単語親密度は単語認知の正確さや速さと極めて強い関係があるとされ，単語を刺激とする心理実験における単語の適切な統制のために調査された．応用として語彙数推定テスト [9] 等に用いられている．

また，天野らは [7] で，新聞記事における単語出現頻度と単語親密度の関係を調査しており，その結果，両者の相関係数は 0.634 で有意な相関がある事がわかっている．同時に，使用頻度は低いが親密度は高い語（例：「たまねぎ」「からあげ」）の存在も確認されたという．

単語の難易度は，出現頻度で近似する手法が用いられる事があるが [2]，そもそも類似の文書が多い CGM（Consumer Generated Media）中では、頻度以外の手がかりも重要である．本研究では，単なる頻度より優れていると天野らが主張している単語親密度を用いれば、文書の難易度をより精密に導けるのではないかと考えた．

3.2 単語親密度とテキスト処理

まず始めに，単語親密度から文章親密度を求める妥当性を評価するために単語親密度辞書と形態素解析の親和性を評価した．

単語親密度を用いて文章の親密度を測る場合，まず文章を形態素に分解する必要があるが，分解した形態素が親密度辞書に含まれない場合，親密度を求める事が出来ない．

そこで，形態素解析エンジンで使用される事の多い IPADIC(注1)と，親密度辞書の分布を調べた．

IPADIC では，活用のある品詞においては，各活用系を展開した数になっている．また，IPADIC は「明るい」と「明かる

(注1)：http://mecab.sourceforge.net/src

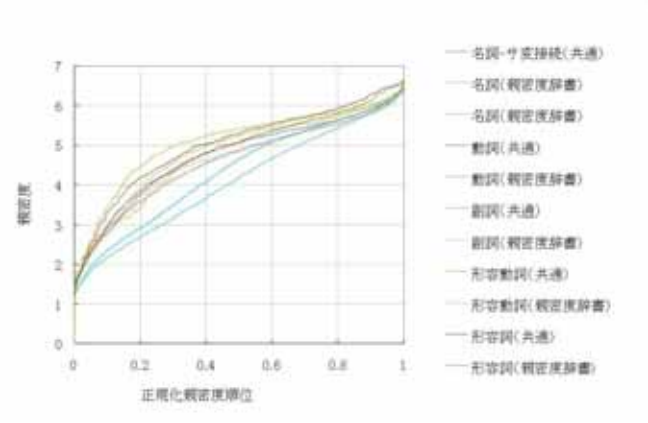

図 1　親密度辞書中の語と IPADIC 共通語の親密度分布

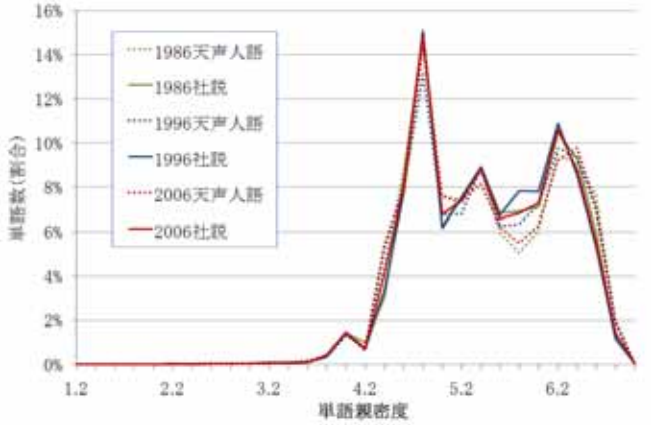

図 2　朝日新聞記事構成単語の親密度分布

い」など異表記を多く含むため，両辞書に共通する語の数は表の程度に留まる．

本来の親密度辞書と，親密度辞書-IPADIC 共通語の分布を比較してみる．図 1 の様に，共通語は各親密度に概ね均等に含まれているのがわかる．

以上から，IPADIC を使った形態素解析で出力される形態素は，概ね親密度辞書によって評価できる事が判った．

以降，IPADIC と形態素解析エンジン MeCab(注2) を用いて，文を単語に分解し，文章の親密度を分析する．

以下の方法で親密度分布を求めた．

（1） 形態素解析エンジン MeCab を使用し，各データの文章を単語に切り分ける．

（2） 単語の表記と品詞名を用いて単語親密度データベースのエントリと照合.（照合できない場合は除去）

（3） 単語親密度を集計し，各データの親密度の分布を求める．

3.3　新聞記事に現れる親密度の差

まず始めに，新聞コーパスを用いて分析を行った．新聞記事は，不特定多数の読者を想定し，文体や語の選定等，細かく統制されているため，読みやすい文章の例として適切である．

朝日新聞中の社説と天声人語のそれぞれにおいて，構成単語の親密度分布を分析した．二つの記事は共に最近のニュースや話題についてのものであるが，異なる角度から執筆されており，一般的に天声人語の方が文章の難易度は低いと考えられる．

新聞記事には朝日新聞記事データベースを使用した[10]～[12]．1986 年，1996 年，2006 年の 10 年おき 3 年分のデータを使用している．データから「天声人語」と「社説」の二種類の記事を抽出し，二種類の記事を 3 年分，計 6 つのデータを比較した．結果を図 2 と図 3 に示す．

図 2 では，記事が含む単語とその親密度を単純に集計した．図 3 は，単語の出現頻度を無視し，記事を構成する単語とその親密度を集計している．

図 2 では，2 種類の記事の間に特に差は見られない．図 3 は，天声人語の親密度分布と比べて社説の親密度分布は低い方に

(注2)：MeCab : Yet Another Part-of-Speech and Morphological Analyzer MeCab, http://mecab.sourceforge.net/

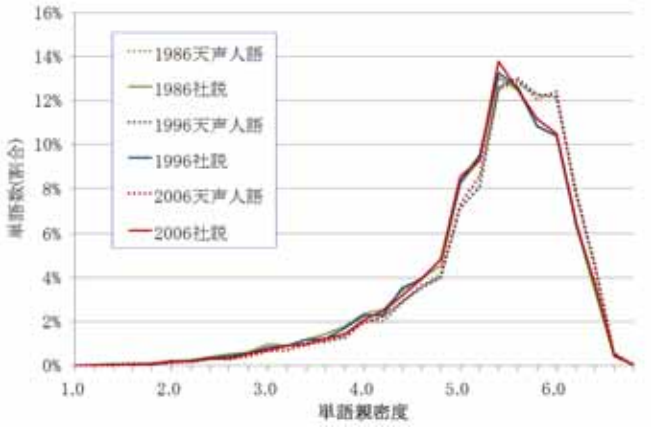

図 3　朝日新聞記事構成単語 (異なり語) の親密度分布

偏っている．

親密度を基に構成単語の分析を行う場合，単語の出現頻度の影響を取り除いた方が良い事がわかる．また，親密度は低いほど「なじみ」が薄い，つまり一般的に難しい単語であるので，グラフは，社説は天声人語よりも難しい言葉を用いて書かれている事を示す．天声人語が，その他の記事と比べて平易に書かれている事は容易に観察できる．この結果は，その事実と概ね合致しているという事ができる．

以上から，異なる種類の記事では，親密度に違いがある事がわかった．

4.　分　　析

本稿では，親密度が CGM テキストにおける文章の評価にも有効であるかを確認するため，以下に示す分析を行う．ここまでに，記事の種類の違いが単語親密度に反映される事を確認した．これを踏まえ，Yahoo!知恵袋で詳細な分析を行う．

4.1　対象データ

次に，Yahoo！知恵袋の記事データを対象に分析を行った．

Yahoo!知恵袋は，利用者が書き込んだ質問に，利用者が回答して書き込む，質問回答サイトである．質問者は，投稿された回答の中から，よいと感じた回答を「ベストアンサー」に選ぶことができる．なお，1 人の利用者は 1 件の質問に対し，1 回

表 2　分析対象カテゴリの記事数

カテゴリ名	PC	恋愛相談
質問	10,086	14,837
ベストアンサー	10,020	14,245
その他の回答	13,989	67,187

だけ回答を書き込むことができる．質問者も含め，一度投稿した内容に追記することはできない仕様になっている．

つまり，ベストアンサーとその質問文の間では，その他の回答に比べて質の高いコミュニケーションが行われている，と考えられる．この仮説をもとに，Yahoo!知恵袋データを用いて分析を行い，ベストアンサーとその他の回答の間の差異・相関とその原因を求めたい．

データは，2004 年 4 月の Yahoo!知恵袋立ち上げ時から 2005 年 10 月までの 19 カ月分のデータを含むが，本稿では 2005 年 9 月の 1 ヵ月分に着目した．これは，立ち上げ時から 19ヶ月の間には利用者が大きく増加しており，利用者のサービスに対する理解も大きく変わっていると考え，全期間のデータを同列に扱う事は出来ないと判断したためである．そのため，利用者数がピークになる 2005 年 9 月のデータを使用している．

4.2　分析方法

本稿では，質問者が選ぶ「ベストアンサー」に着目する．質問回答サイトにおける一連のコミュニケーションでは，投稿されたテキストの中から読み取れる範囲という制約はあるものの，質問者が抱える問題を正確に理解することや，質問者の感情をくみ取ることなどが回答者に求められる.「ベストアンサー」に選ばれた回答およびその投稿者は，質問者を満足させることができたといえる．また，ベストアンサーとその他の回答には，何らかの差があると考えられる．質問とベストアンサー，ベストアンサーとその他の回答，質問とその他の回答について，テキスト中に出現する語の親密度および出現頻度を比較し，親密度が有効な評価指標となることを確認する．分析手順は，新聞と同様，MeCab を用いて単語に分割し，単語の表記（漢字・ひらがな・カタカナで表記される）と品詞名を用いて単語親密度データベースのエントリと照合し，一致するエントリが存在する単語のみを分析に用いる．

4.3　分析結果

対象とする記事をカテゴリで絞り込んだ．先行研究 [13] により，傾向の違いが指摘されている 2 大カテゴリの「パソコン，周辺機器」（以下，PC カテゴリ）と「恋愛相談，人間関係の悩み」（以下，恋愛カテゴリ）を対象とした．それぞれの記事数は表 2 の通りである．

この記事を対象に，質問・ベストアンサーの間の相関を分析する．

4.3.1　記事ごとの平均親密度

記事に含まれる全ての単語を用いて，記事ごとに単語親密度の平均を求めた．カテゴリ別に，質問と，その質問に付随するベストアンサーにおける，平均親密度を比較した結果を図 4，図 5 に示す質問，ベストアンサーともに，親密度 6 をピークとした分布となっており，特に相関は見られなかった．相関係数は，PC カテゴリでは 0.10，恋愛カテゴリでは 0.14 となった．

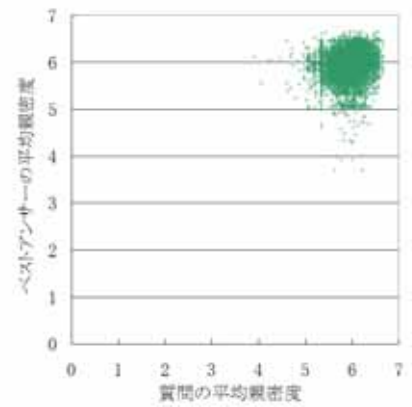

図 4　PC カテゴリにおける質問とベストアンサーの平均親密度分布

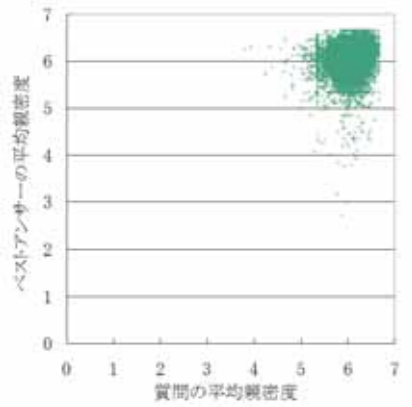

図 5　恋愛相談カテゴリにおける質問とベストアンサーの平均親密度分布

4.3.2　品詞別の出現頻度

親密度を品詞別に測ることで，テキストの性質の違いを検出できるか確認するため，品詞ごとの各単語の出現頻度を調べた．未知語の抽出処理 [3] [4] や評価表現の抽出 [5] において利用されることの多い品詞を参考に，一定以上の単語数を有し，親密度の異なる語の使い分けが容易と考えられ，かつ主題語になりやすい語の少ない品詞として，サ変名詞 (サ変動詞「する」に接続可能な名詞)，副詞，形容動詞（名詞の形容動詞語幹），形容詞に着目する．各カテゴリ，品詞，投稿区分における頻出単語の例を表 3, 表 4 に示す．

各カテゴリにおいて，これらの 4 品詞における出現頻度の順位を正規化し，質問 (Q)，ベストアンサー (BA)，その他の回答 (NA) の各回答区分間で比較した結果を，表 5，表 6 に示す．表中の $R_{(q)}$，$R_{(ba)}$，$R_{(na)}$ は，Q， BA，NA それぞれの投稿区分における出現頻度順位を正規化した値である．

PC カテゴリでは、形容動詞以外の 3 品詞のベストアンサー-その他の回答が 0.8 以上で非常に強い相関が見られる．つまり、ベストアンサーで頻出する語は、その他の回答でも頻出することを意味する．ただし、サ変名詞については質問文-ベストアンサー、質問文-その他の回答でも 0.7 以上であり相関は強く、サ変名詞には記事の種類を問わず同様に使われる主題語や一般語

表 3　PC カテゴリにおける頻出単語（品詞別）

品詞	サ変名詞			副詞			形容動詞			形容詞		
順位	Q	BA	NA	Q	BA	NA	Q	BA	NA	Q	BA	NA
1	表示	設定	設定	どう	どう	どう	簡単	簡単	簡単	詳しい	無い	無い
2	質問	表示	表示	よろしく	まず	そう	自動的	自動的	残念	新しい	多い	多い
3	使用	起動	使用	なぜ	そう	まず	ダイレクト	残念	自動的	無い	新しい	高い
4	設定	チェック	質問	もう	もし	もし	フル	フル	フル	悪い	高い	新しい
5	メール	削除	ドライブ	どうして	多分	もう	アカデミック	私的	厳密	古い	安い	安い
6	起動	選択	起動	突然	かなり	かなり	膨大	シンプル	的確	多い	古い	悪い
7	削除	使用	削除	もし	どうぞ	少し	明らか	明らか	私的	安い	大きい	古い
8	入力	入力	入力	全く	特に	ちょっと	シンプル	厳密	明らか	おかしい	悪い	大きい
9	購入	ドライブ	コピー	やはり	もう	特に	決定的	素直	シンプル	欲しい	早い	早い
10	保存	保存	回答	ちゃんと	少し	ちゃんと	残念	ダイレクト	素直	高い	遅い	欲しい

表 4　恋愛カテゴリにおける頻出単語（品詞別）

品詞	サ変名詞			副詞			形容動詞			形容詞		
順位	Q	BA	NA	Q	BA	NA	Q	BA	NA	Q	BA	NA
1	質問	結婚	結婚	どう	どう	どう	素直	素直	素直	悪い	悪い	悪い
2	メール	関係	関係	もう	そう	そう	簡単	素敵	簡単	多い	多い	無い
3	結婚	メール	仕事	そう	もう	もう	素敵	簡単	素敵	欲しい	無い	多い
4	仕事	恋愛	メール	やっぱり	本当に	本当に	明らか	残念	残念	無い	欲しい	欲しい
5	話	仕事	質問	本当に	ちょっと	もっと	かわいそう	立派	かわいそう	嬉しい	楽しい	若い
6	関係	話	浮気	とても	きっと	ちょっと	残念	かわいそう	立派	おかしい	辛い	楽しい
7	電話	質問	話	ちょっと	もっと	きっと	些細	気楽	気楽	辛い	若い	辛い
8	恋愛	浮気	恋愛	少し	少し	少し	性的	些細	大嫌い	寂しい	嬉しい	嬉しい
9	告白	経験	不倫	どうして	もし	もし	大嫌い	大嫌い	明らか	忙しい	新しい	高い
10	浮気	生活	離婚	やはり	ちゃんと	ちゃんと	気軽	新鮮	些細	楽しい	高い	新しい

表 5　PC カテゴリにおける投稿区分ごとの出現頻度順位の相関係数（品詞別）

品詞	$R_{(q)}, R_{(ba)}$	$R_{(q)}, R_{(na)}$	$R_{(ba)}, R_{(na)}$
サ変名詞	0.728	0.728	0.833
副詞	0.603	0.672	0.870
形容動詞	0.404	0.337	0.757
形容詞	0.701	0.649	0.846

表 6　恋愛カテゴリにおける投稿区分ごとの出現頻度順位の相関係数（品詞別）

品詞	$R_{(q)}, R_{(ba)}$	$R_{(q)}, R_{(na)}$	$R_{(ba)}, R_{(na)}$
サ変名詞	0.764	0.771	0.820
副詞	0.796	0.797	0.905
形容動詞	0.651	0.669	0.795
形容詞	0.873	0.866	0.914

が含まれていると見られる．形容動詞では、質問文-ベストアンサー、質問文-その他の回答の相関が 0.4 以下であり相関は弱く、質問と回答とで異なる語が使われている度合いが大きいことがわかった．また、副詞も質問文-ベストアンサー、質問文-その他の回答は 0.6 程度であり、ベストアンサー-その他の回答よりは相関が弱かった．恋愛相談カテゴリでは、サ変名詞については PC と同様であるが、副詞と形容詞の相関が非常に強かった．形容動詞についても、PC よりは相関が高めだった．

4.3.3　出現頻度と親密度

各カテゴリ，各品詞，各投稿区分における，単語の出現頻度順位と親密度の相関を調べた結果を表 7，表 6 に示す．二つの相関が高い場合，なるべく親密度の高い単語を使おうという意図が伺える．

PC カテゴリでは、副詞において、質問文とベストアンサーの相関係数がやや高かった．また、形容動詞と形容詞においては、ベストアンサーとの相関が高かった．つまり、質問では親密度の高い平易な副詞が多く用いられる傾向がある．また、ベストアンサーはその他の回答よりも、親密度の高い副詞、形容動詞、形容詞が多く使われる傾向にある．

表 7　PC カテゴリにおける親密度と投稿区分ごとの出現頻度順位の相関係数（品詞別）

品詞	$wf,\ R_{(q)}$	$wf,\ R_{(ba)}$	$wf,\ R_{(na)}$
サ変名詞	0.301	0.257	0.311
副詞	0.437	0.423	0.397
形容動詞	0.209	0.436	0.207
形容詞	0.263	0.315	0.227

表 8　恋愛カテゴリにおける親密度と投稿区分ごとの出現頻度順位の相関係数（品詞別）

品詞	$wf,\ R_{(q)}$	$wf,\ R_{(ba)}$	$wf,\ R_{(na)}$
サ変名詞	0.486	0.464	0.422
副詞	0.507	0.446	0.463
形容動詞	0.401	0.476	0.314
形容詞	0.562	0.545	0.510

サ変名詞については、質問文、ベストアンサー、その他の回答のいずれとも相関は弱かった．サ変名詞には「表示」「削除」「設定」など、PC カテゴリにおいて主題語となる語が多く含まれていることから、投稿区分に関係なく、また親密度に関係なく、頻出していることがわかる．

恋愛相談カテゴリにおいては、いずれの品詞・投稿区分においても、PC カテゴリより強い相関が見られた．特に形容詞では 0.5 以上であり、やや強い相関が見られた．恋愛相談カテゴリでは、感情や印象を表現するため、そもそも形容詞が多用される（異なり語数が PC より多い）．その上で、さらに親密度の高い形容詞ほど頻出する傾向があることを意味している．また，PC カテゴリほど顕著ではないが、副詞においてはベストアンサー、その他の回答よりも質問文の相関が高くなっている．

4.4 考　　察

記事ごとに，全品詞の単語親密度を単純に平均する手法では，テキストの特徴が把握できないことがわかった．そこで，サ変名詞，副詞，形容動詞，形容詞に着目したが，品詞やカテゴリによっては，主題語や一般語が多く含まれることがわかった．一般的な情報検索と同様に，内容によらず出現頻度の高い一般語を除去する前処理が必要であるといえる．

PC カテゴリにおいては,「印刷」「設定」「削除」などの語が頻出していた．これらの語は，必ずしも質問回答サイトの全体を通じて一般語であるわけではないが，当該カテゴリ内においては一般語であるといえる．このような語については，何らかの方法で取り除いたり，必要に応じて残したりできることが望ましい．

また，形容動詞として「アカデミック」「プロフェッショナル」などの語が頻出していたが，これはソフトウェアの「アカデミック版」「プロフェッショナル版」から切り出された語であり，形容動詞として使われているわけではない．同様に,「設定変更」からは「設定」「変更」がそれぞれサ変名詞として切り出されるが，サ変動詞「する」に接続しない用法で使われたサ変名詞を除くなどの処理を追加することも検討する必要がある．

今後の課題として，単純な平均や出現頻度ではなく,「わかりやすく書きたい」あるいは「格調高く書きたい」といった書き手の意図や，同じ単語であっても，漢字や送り仮名など表記の違いによって異なると考えられる読み手の印象などを含めて考慮できる指標の構築が必要と考えられる．

5. ま と め

本稿では，テキストの印象を決定する重要な要因としてテキストの難易度を取り上げ，単語の難易度からテキスト全体の難易度を推定する手法を検討した．単語の難易度として語彙の認知度を網羅した単語親密度を紹介し，形態素解析エンジンとの親和性を評価した後，新聞記事・Yahoo!知恵袋の記事を対象に分析を行った．分析の結果，記事が含む単語が持つ単語親密度を平均する単純な手法では，テキストの特徴は把握できない事がわかった．今後，一般語や主題語の取扱い等を考慮して文の親密度を計算する手法を考える必要がある．

謝　　辞

本研究の実装・評価に際し，大学共同利用機関法人 国立情報学研究所から提供を受けた，Yahoo!知恵袋のデータを利用している．ここに記して謝意を示す．また，本研究の一部は,「平成 20 年度筑波大学図書館情報メディア研究科プロジェクト研究」の支援を受けて行われました．

文　　献

[1] Mehrabian, A, Silent messages, Wadsworth, Belmont, California, 1971.
[2] 西原陽子，砂山渡，谷内田正彦, Web ページの難易度と学習順序に基づく情報理解支援システム, 電子情報通信学会論文誌, Vol.J89-D, No.9, pp.1963-1975, 2006.
[3] 福島健一，鍜治伸裕，喜連川優, 機械学習を用いたカタカナ用言の獲得, 言語処理学会第 13 回年次大会発表論文集, pp.815-818, 2007.
[4] 桑江常則，佐藤理史，藤田篤, 後続ひらがな列に基づく語の活用型推定. 情報処理学会研究報告, NL-186-2, pp.7-12, 2008.
[5] 峠泰成，山本和英, 手がかり語自動取得による Web 掲示板からの評価文抽出, 言語処理学会第 10 回年次大会, pp.107-110, 2004.
[6] 天野成昭，近藤公久, 日本語の語彙特性 1　単語親密度, 三省堂, 東京, 1999.
[7] 天野成昭，笠原要，近藤公久, 日本語の語彙特性 7 単語親密度増補, 三省堂, 東京, 2008.
[8] 天野成昭，笠原要，近藤公久, 日本語の語彙特性 9 単語親密度増補, 三省堂，東京，2008.
[9] 天野成昭，近藤公久，片岡良治, 単語親密度を利用した語彙数推定：インターネットによる大規模調査, 日本認知科学会第 22 回大会予稿集, pp.58-59, 2005.
[10] 朝日新聞記事データ集 1986/1987 年版, 日外アソシエーツ.
[11] 朝日新聞記事データ集 1996 年版, 日外アソシエーツ.
[12] 朝日新聞記事データ集 2006 年版, 日外アソシエーツ.
[13] 島田諭，福原知宏，佐藤哲司, 質問回答サイトにおける利用者行動に基づく記事の関連付け手法の検討, DEWS2008, B6-5, 2008