文章の判りやすさの定量的な指標について

石原 亜紗希 †　　中山 伸一 ‡　　真栄城 哲也 ‡

† 筑波大学 図書館情報専門学群　　‡ 筑波大学 図書館情報メディア研究科

1 はじめに

本研究では，文章の判り易さを理解度と通読性の側面から捉え，係り受けの距離，構文の深さ，文節毎の文字数のゆらぎの3つの数量的な指標を用いて文章を解析する．文章の判りやすさについては様々な研究が発表されている [1, 3]．判り易さには様々な要因があり，文章の判り易さを向上させる方法をまとめた本も多く出版されている ([2] 他)．

しかし，改善手法が完全に人手によること，文章の内容への依存度が高いこと，改善する人の能力に大きく依存すること等から，これらの本に記載されている内容の効果については，普遍性に欠ける側面がある．一方，判り易さを自動で向上，あるいは機械的に改善する方法は確立されておらず，人手による作業が必要である．本研究は，文章の判り易さを機械的に改善する際に必要な数量化の手法の構築を目的としている．

文章の判り易さには様々な側面があるが，本研究では文章の理解度と通読性を扱う．理解度とは，文脈や文節間の関係等，文章の内容を違えることなく理解できる度合いとし，また，通読性とは，音読や黙読する際に，文章を間えることなく読める度合い，とする．本論文は，理解度および通読性の良さに関わる構文構造の特徴の数量化と評価について述べる．

2 方法

判り易さの指標として，主に経験則に基づくと思われる指標が多数提唱されている．従って，文章の書き方について書かれた計20冊の本から文章の判り易さに影響する要因を抽出した．これらから得られた417項目を分類し，23のグループに属する合計98項目を得た．さらに，数量化や自動化の可能性に基づき，判り易さへの影響を解析するために以下の3つの要因を選択した．(A) 文章を句読点単位で区切った際の文字数のゆらぎ，(B) 係り受け距離，(C) 構文木の深さ．以下，要因 A, B, C とする．

要因 A (ゆらぎ) については，2つの内容の異なる文章それぞれに論理構造の異なる文章のペアを用意し，サンプル数32で FFT (高速フーリエ変換) により解析する (表1の文章 α と β)．周波数成分の分布によって句読点間の文字数のゆらぎが計測できる．他の文章についても同様の解析が可能だが，これらの4つの文章は視線移動の計測によって文章の読み易さに関する詳細なデータが計測済である [4]．[4] では，文章を論理構造のブロックに分割し，ブロック間の論理関係と包含 (階層) 関係によって構造化している．この論理的な階層構造と，文章を読む際の視点移動データを解析した結果，注視回数および注視時間に有意な差があり，どちらの値も判りやすい文章の場合が長くなることが判った．従って，読む際のリズムに関わる文章の定量的な特性について分析する．

係り受け距離 (要因 B) と構文木の深さ (要因 C) については，以下のような被験者実験により検証する．異なる内容および文字数の文章を6個用意し，被験者実験および数値解析により選択した3つの要因の妥当性を検証する．文章毎に，係り受け距離を変更した文章を生成した．6つの文章それぞれに，係り受け距離が異なる文章が2通りあるため，ペア毎に被験者に読んでもらい，判り易いと判断した文章を選んでもらう．全ての文章 (計12) を提示するが，被験者毎にペアの提示順序やペア間の順序を変更する．

用意した文章の文字数は表1の通りである．文章 I と II は，視線移動についてのデータがある文章 α と β の一部をそれぞれ抜粋したものである．係り受け距離の変動範囲を最小限にするため，要因 B と C の解析には短い文章を用意した．一方，要因 A であるゆらぎについては，ある程度以上の長さの文章が必要なため，3,000文字以上の比較的長い文章を用意した (表1)．

[1] Some Quantitative Indicators of Text Understandability
[2] Asaki Ishihara, University of Tsukuba
[2] Shin-ichi Nakayama, University of Tsukuba
[2] Tetsuya Maeshiro, University of Tsukuba

表 1: 被験者実験で使用した文章の文字数．a と b は，文章 α と β については論理構造が異なるペア，文章 I〜VI については係り受け距離が異なる文章のペアである．

	a	b
文章 α	3,195	3,096
文章 β	3,371	3,588
文章 I	311	306
文章 II	398	397
文章 III	160	160
文章 IV	100	100
文章 V	44	44
文章 VI	56	56

文章 I〜VI は，文章全体の係り受け距離を長くした文章と，構文構造を表す構文木の深さを増加した文章をを用意する (表 1 の文章 a と b)．後者は，文節の連なりを長くする操作を含む (表 1 の文章 a と b)．なお，構文解析には KNP [1] を用いる．

3 結果および考察

要因 A (ゆらぎ) については，文章 α と β を FFT で解析した．その結果，文章 α と β 共に論理構造の判りやすさに関連して，強度の高い周波数成分およびピーク周波数の数と帯域に差があった．周波数成分がベキ分布の場合，1/f ゆらぎと称される人間に心地良いリズムだと考えられている．しかし，本研究では高周波帯域が高い結果も出ているため，1/f ゆらぎとの関連は明確ではない．別の原理が働いている可能性や無関係である可能性があり，より詳細な解析が必要である．

要因 B と C については，21 人の大学生を対象に行った．その結果，係り受け距離が長い文章と構文木の深い文章のペアの内，分かり易いと選択された文章は文章に依存した (表 2)．選択された文章は 100%：0% から 23.5%：76.5% の範囲で変化し，選択が半分ずつの文章もあった．これらの結果から，係り受け距離の長い方がより分かり易いと判断される傾向が強いが，必ずしも成立せず，他の要因の影響があると考えられる．

従って，文章 I〜VI それぞれの文章 (a) と (b) を主語と述語に着目して解析したところ，主語と述語間の距離が関与していることが示唆された．文節毎の係り受け距離を同一文章のペア (a) と (b) で比較すると，係り受け距離の増減の度合いと，(a) が分かり易いと選択された割合に関連性が見られた．

表 2: 同じ文章を元に，係り受け距離が長い文章と構文木が深い文章を生成し，それぞれが分かり易いと選択された割合．

	係り受け距離が長い	構文木が深い
文章 I	76.5%	23.5%
文章 II	57.1%	42.9%
文章 III	76.2%	23.8%
文章 IV	100%	0%
文章 V	47.6%	52.4%
文章 VI	23.5%	76.5%

本研究では，文章 I〜VI から係り受け距離と構文木の深さを変動させて文章を生成するが，両者は関連するパラメータである．ここで用いた 12 個 (6 ペア) の文章は，両パラメータを同時に変化させるように文章生成した．一般的に，一方が増加すれば他方は減少するが，例外も存在するため，係り受け距離と構文木の深さをどちらも判り易さの要因とした．両パラメータの値の関係が反比例であれば，本研究で用いた文章の判り易さの要因 (B) と (C) を 1 つに集約できるが，さらなる検討が必要である．また，構文解析についても検討を要する．

参考文献

[1] 秋野喜代美,「文章理解の心理学：認知，発達，教育の広がりのなかで」，北大路書房，2001

[2] 藤沢晃治.「分かりやすい」文章の技術．講談社．2004

[3] 高橋義文ほか．計算機マニュアルの分かりやすさの定量的評価方法，情報処理学会論文誌 32(4), 460–469, 1991

[4] 鈴木麻衣子，中山伸一，Haakon Lund, 真栄城哲也,「文章の読み方および論理構造に関する判りやすさの計測」，情報処理学会第 70 階全国大会論文集，4:91–92, 2008

[1] http://nlp.kuee.kyoto-u.ac.jp/nl-resource/knp.html