福井県教育総合研究所 紀要 第 127 号
特別寄稿

単語親密度と語彙数推定

－育児・教育支援への応用を目指して－

NTT コミュニケーション科学基礎研究所

藤田早苗

同じテキストでも、難しいと感じるか易しいと感じるかは、読み手によって異なる。これは、読み手の知識量が異なるためであり、もし、読み手にとってちょうど読めるくらいの、あるいは少し頑張れば読めるくらいの絵本や本、英文を薦めることができれば、無理なく読み手の知識を増やしていけるかもしれない。しかし、「ちょうど読めるくらい」や「少し頑張れば読めるくらい」を推薦するのは簡単ではない。文（テキスト）側の難易度と、人側の知識量の両方を適切に推定する必要があり、さらにそれらを組み合わせて推薦する必要があるからである。NTT では、文（テキスト）の難しさを自動的に推定するための難易度推定や、読み手側の知識量（語彙数）の推定方法、推薦方法などの研究に取り組んでいる。その中から本稿では主に、人の知識量（語彙数）の推定方法を紹介し、推定結果を育児・教育支援に活かす取り組みも最後に紹介する。

＜キーワード＞　難易度推定　単語親密度　語彙数　教科書　絵本

Ⅰ　人の語彙数の調査・推定方法

NTT では 20 年以上前から、様々な年齢層の人の語彙数の調査や推定に取り組んできている。

乳幼児の場合、語彙数自体は多くないので、理解/発話できるすべての語彙の調査に取り組んでいる。そのため我々は、1500 組以上の親子モニターの皆さんにご協力をいただき、子どもがいつごろどのような語を覚えるか、発話できるかというデータを蓄積し、「幼児語彙発達データベース」を構築した。

しかし小学生以上となると、知っているすべての語彙を調査することは困難である。そこで、提示した少数の語を知っているか回答してもらうことにより、語彙数を推定する方法をとっている。提示する語は多いほど正確な推定ができるが、数十語でも推定は可能である。この方法では、ある語を知っていると回答したときに、何語知っていると仮定するかがポイントとなる。例えば「銀行」と「地歩」だと、「地歩」という語を知っている人の方が少ないだろう。そのため、「銀行」だけを知っている人より「地歩」も知っている人のほうが、より多くの語を知っていると仮定できる。では「地歩」を知っていれば「何語くらい知っている」と推定できるのだろうか。

その推定の根拠となるのが、語のなじみ深さを数値化した「単語親密度」（以下、親密度）である。親密度は、1-7 の間の数値で表され、値が大きいほど多くの人にとってなじみ深い語であることを示している。NTT では 20 年以上前に約 7 万 7 千語の親密度を調査（平成版親密度）、それを元に語彙数推定テストを公開し、広く利用されてきた。しかし、調査から 20 年以上が経過したことから、新しい語の追加と再調査を実施、約 16 万 3 千語からなる、より大規模な「令和版単語親密度データベース」を構築した。図 1 に各親密度の語の度数分布をしめす。（本データベースは NTT 印刷(株)から有償で提供している。https://www.nttprint.com/lexicon-db/）[2]。

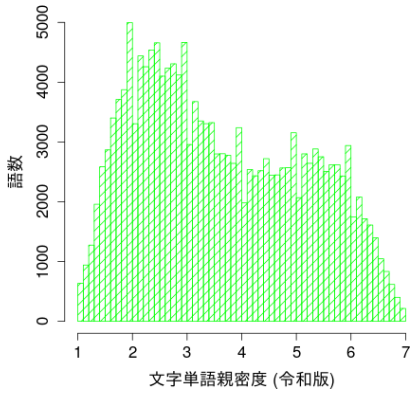

図 1　単語親密度の度数分布

本データベースによると、「銀行」は親密度が 6.667、「地歩」は 2.169 である。親密度が 6.667 以上の語は約 1,400 語だけだが、2.169 以上の語は 10 万語以上である。ただし、知っている語と知らない語の境界付近では、知っているかどうかはばらつく可能性がある。例えば、「地歩」を知っていても、近い親密度の「四十物」という語は知らないかもしれない。

そこで、「知っているかどうか」の回答結果から各親密度の語を知っている確率を計算する。図２はある中学生の回答結果である。図２の生徒の場合、知っている確率が 50%となる親密度（以降、50%獲得親密度）は 4.1 である。これを語彙数に換算する場合には、確率値の積算などを用いる。ただし、算出した語彙数はあくまで推定であり、絶対的な正解ではない点は注意する必要がある。

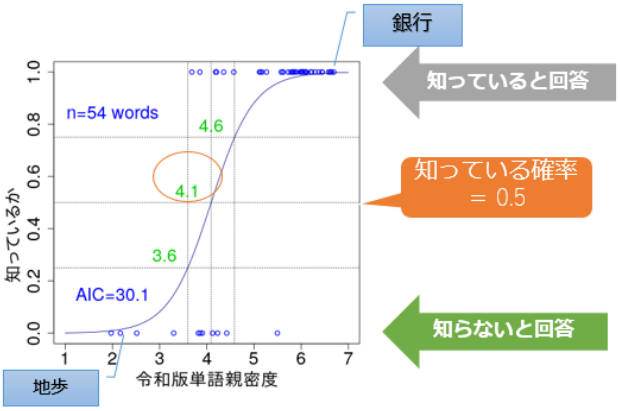

図２　単語親密度と知っている確率の算出例

NTT では、少数の語のチェックから回答者の語彙数を簡単に推定できる「令和版語彙数推定テスト」を作成し、2020 年６月から Web 版として公開しており、2022 年２月までに７万６千人以上の方にご利用いただいている（https://www.kecl.ntt.co.jp/icl/lirg/resources/goitokusei/vocabulary_test/php/login.php）[5]。

Ⅱ　語彙数推定調査

単語親密度に基づく語彙数推定には、様々なバリエーションのテストを簡単に作成でき、テストを受ける側の負担も少ないという利点がある。我々は、Web 版以外にも何通りかのテストを作成、調査に利用してきている。本稿ではその調査結果からいくつかを紹介したい。

１　学年・年代による変化

2018 年度から 2020 年度にかけ、公立の小学６年生～高校生と大人を含む約 14,000 人を対象に行った調査を紹介する。図３には 50%獲得親密度の分布を、図４には語彙数に換算した結果を示す。図３，４からは、各学年・年代ごとに語彙数が増えていく様子が見て取れる。小中学生は急激に語彙数が増えているが、成人でも緩やかに語彙数は増えている。高校生のみ語彙数の減少が見られるが、高校の場合は調査対象の学校によって結果が大きく変わること、調査人数が少ないことが要因として考えられる。本調査結果からは、同じ学年・年代でもばらつきが大きいこともよくわかり、支援が必要な生徒を見つけることにも役立つと考えられる[3]。

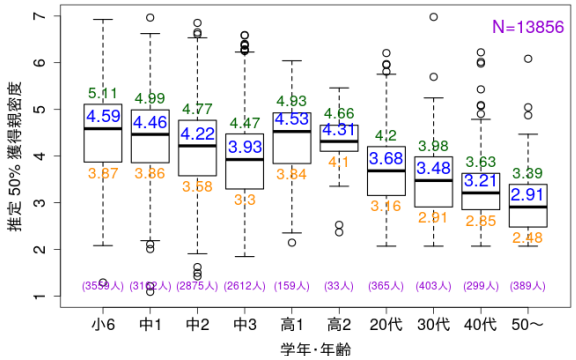

図３　各学年・年代における 50%獲得親密度の分布

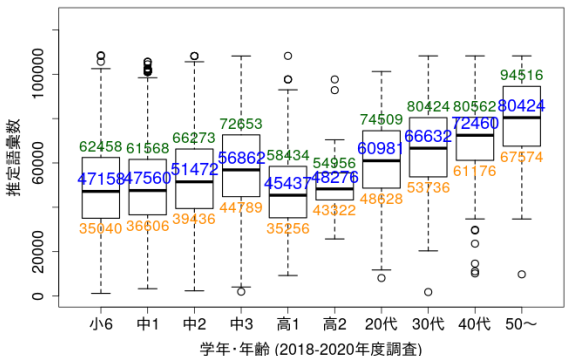

図４　各学年・年代における推定語彙数の分布

2　単語親密度と獲得割合

前節では一人一人の推定結果の分布を紹介したが、本節では、単語親密度とその親密度の語の獲得割合（その語を知っている人の割合）との関係を紹介する。両者の関係を、各学年・年代でモデル化したのが図５である。どの学年・年代でも親密度が高い語ほど、知っている人の割合は高くなる傾向があり、年齢が上がれば上がるほど、この傾向は顕著になる。一方、成人にくらべて、小学生や中学生では、児童・生徒ごとに個人差が大きく、比較的親密度の高い語であっても、知っているかどうかはばらつきが大きい。

そもそも親密度は、成人（20-30 代）での調査結果から算出されており、子どもにとっての親密度とは必ずしも一致しない可能性がある。しかしながら、年齢とともに収束していく様子から、児童・生徒がこれから重点的に獲得するだろう語彙、あるいは個人ごとに獲得した方がよい語彙を見つける手掛かりとして、親密度を利用できるのではないかと考えている。

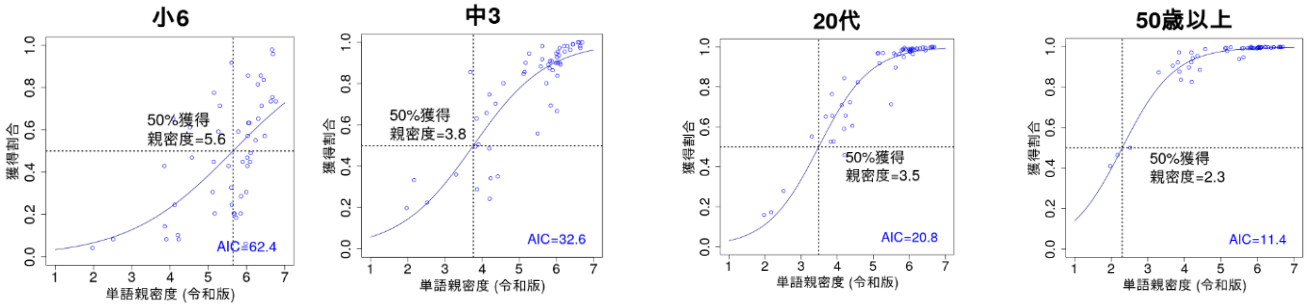

図５　各学年・年代における語彙獲得状況と単語親密度

3　語彙数のより正確・継続的な調査

ここまで紹介してきたように、親密度に基づく語彙数推定は簡便で拡張性が高い。しかし一方で、「知っているかどうか」の判断基準が人によって異なるという問題がある。そこで、正誤問題と組み合わせたより正確な語彙数推定の継続的調査にも取り組んでいる。

具体的には、まず第１段階で従来どおり「知っているかどうか」を回答してもらい、第２段階では「知っている」と回答した語のうち親密度の低い方から５語を取り出して４択問題を出題する。この４択問題でも正解した語は「知っている」、間違えた語は「知らない」として語彙数などを推定する。

我々は大阪の公立小学校で、小学４年生から６年生の児童約 360 人を対象に、2021 年２月と 2021 年９月に調査を実施した。同じ児童での 50%獲得親密度（推定語彙数）の変化を調べると、50%獲得親密度は平均-0.33、語彙数では約 2300 語増加していた。本調査は継続しており、現在（2022 年２月）も実施中である。

Ⅱ　教科書中の語彙と単語親密度

ここまで、人の語彙数の推定方法と調査結果を紹介してきた。では、どの程度の親密度の語までわかればどの程度の文章（テキスト）が読めるのだろうか。

テキストの難しさには、使われている語彙の難しさ、文構造の難しさの両方が影響する。そのため我々は、テキストの難易度推定を行う場合、幼児語彙発達データベースや親密度などの語彙の難しさを反映する特徴量と、文の長さ等の文構造の難しさを反映する特徴量を用いて難易度推定を行っている[1]。それにより、絵本の難易度推定なども行っているが、本稿では、国語教科書中の語の親密度の分布を紹介したい。

我々は、小学１年生～中学３年生用の、シェアトップ３社（光村書店、教育出版、東京書籍）の 2014 年検定の国語教科書をすべて書き起こし、そのテキスト中に出てくる語の親密度を調査した。図６に、小１，小６，中３ の光村書店の国語教科書に出現する語の親密度の分布を示す。縦軸は異なり語数であり、何度も出てくる語でも同じ語は１語として数えている。また、固有名詞など、親密度が付与されていない語は除いている。

図６からわかるように、学年が上がると異なり語数は増えるが、山のピークはいずれの学年でも親密度６以上と高い。令和版単語親密度データベース(図１)と比較すると、山の形は明らかに異なっている。令和版単語親密度データベースには親密度２～３の語も多く存在するが、少なくとも国語教科書を読むのに必要な語は親密度の高い語が多く、まずはこうした親密度の高い語を取りこぼしなくわかるようにしていけば教科書を無理なく読めるようになっていくだろう。

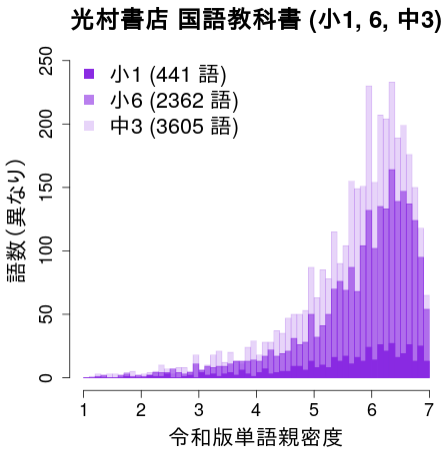

図６　国語教科書中の語の単語親密分布

Ⅲ　育児・教育支援への応用

本稿では、人の語彙数推定方法やテキストの難易度推定方法等について紹介してきたが、こうした研究を育児・教育支援に活かす取り組みも始めている。

我々はまず、一人一人の子どもにあった読みやすさで興味のありそうな内容の絵本を探すためのシステム「ぴたりえ」を作成した(紹介動画: https://www.youtube.com/watch?v=xhyKKJIWtkE)。ぴたりえは、(株)NTT データ九州から図書館向けに提供しており、福井県立図書館には 2019 年にファーストユーザーとして導入していただいている（そんなわけで、福井県には勝手にご縁を感じている。いつかは恐竜好きの息子と一緒に、恐竜博物館への再訪と県立図書館訪問を叶えたい)。

ぴたりえでは、保護者や司書といった大人が絵本を選んであげることを想定していたが、小さな子どもでも自分で絵本を選べるようにと作成したのが「ぴたりえタッチ」である。ぴたりえタッチでは、ロボットからの質問に答えると絵本を推薦してくれる(図７)。現在、兵庫県西宮市や沖縄県恩納村の図書館で実証実験中である(西宮市による紹介動画(ぴたりえタッチは 6:25 頃から): https://www.youtube.com/watch?v=22HYt8P4wis)。

図７　ロボットによる絵本推薦

さらに我々は、日本語だけでなく、英語の語彙数推定と難易度推定の研究も進めている。その一環として、英語の語彙数を推定し、語彙数にあった英語のテキスト（絵本や記事など）や問題の推薦を繰り返すことによる、英語の学習支援にも取り組んでいる（図８，[4])。

図８ 英語の語彙数推定と絵本・問題推薦

我々は今後も、日本語でも英語でも、幼児でも小中高校生でも、大人に対しても、エビデンスを積み重ねながら、一人ひとりにあった育児・教育支援の実現を目指していきたい。

関連発表

[1] 藤田早苗, 小林哲生, 南泰浩, 杉山弘晃, ″幼児を対象としたテキストの対象年齢推定方法,″ 認知科学, Vol. 22, No. 4, pp. 604-620, 2015.

[2] 藤田早苗, 小林哲生, “単語親密度の再調査と過去のデータとの比較,” 言語処理学会第 26 回年次大会 (NLP), 2020.

[3] 藤田早苗, 小林哲生, 山田武士, 菅原真悟, 新井庭子, 新井紀子, “小・中・高校生の語彙数調査および単語親密度との関係分析,” 言語処理学会第 26 回年次大会 (NLP), 2020.

[4] 藤田早苗, 服部正嗣, 小林哲生, 納谷太, “日本人初学者の語彙数推定方法の検討,” 2020 年度人工知能学会全国大会 (JSAI), 2020.

[5] 藤田早苗, 小林哲生, “令和版単語親密度に基づく大規模語彙数推定調査 ～Web 公開版の利用ログ分析～” 2022 年度人工知能学会全国大会 (JSAI), 2022. (to appear)