言葉の難しさを測る──テキストの難易度と人の語彙数の推定

文章を読むときに，難しいと感じるか，易しいと感じるかは，文章の難易度と読み手自身の知識量の両方に依存します．文章（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と，読み手側の知識量（語彙数）の推定が，簡単に，高精度にできれば，ちょうど良い難しさのテキストを推薦することができ，学習支援につなげることも可能です．本稿では，私たちが取り組んでいるテキストの難易度推定方法と語彙数推定方法を紹介します．

藤田（ふじた）　早苗（さなえ）
NTTコミュニケーション科学基礎研究所

難しさを測る意味

文字を覚えたての子どもが自分で読むつもりで選んだ絵本が読めず，読んであげることになったことはありませんか．中学１年生のときとても苦労して読んだ英文が，大学生になるころにはとても簡単に感じられたことはありませんか．同じ文を読もうとしても，難しいと感じるか易しいと感じるかは，読み手の知識量に依存します．

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ません．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」を判断するのは簡単ではありません．人の知識量と文（テキスト）の難易度の両方を適切に推定する必要があるからです．

本稿では，この両方の推定方法の研究と，推定を支える言語資源の構築について紹介します．

人の語彙数を測る

人に必要な知識の１つとして語彙の知識が挙げられます．NTTでは20年以上前から，さまざまな年代の人の語彙数の調査や推定に取り組んできました．

幼児を対象とした調査では，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を調査することも不可能ではありません．実際私たちは，1500組以上の親子モニターの皆様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てきました．

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す．そこで，提示した語を知っているか回答してもらうことにより，語彙数を推定します．提示

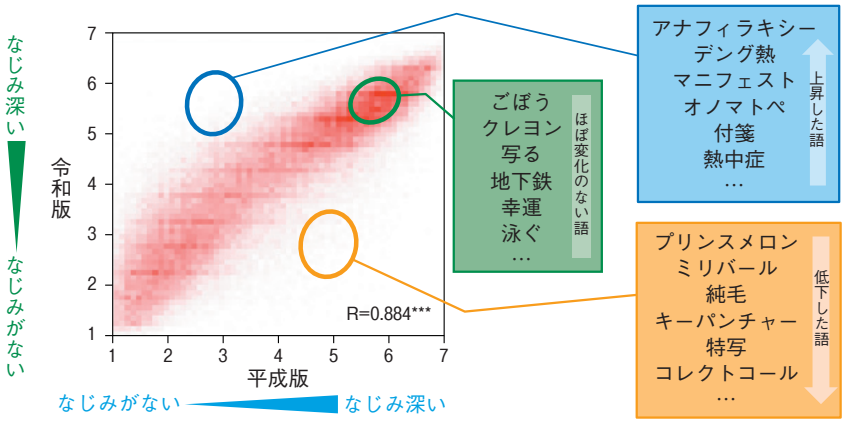

図１　平成版から令和版への単語親密度の経年変化

する語は多ければ多いほど正確な推定ができますが，数十語でも推定可能です．

この推定方法では，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ります．例えば「しょっぱい」と「検潮儀」だと，「検潮儀」のほうが知っている人は少ないでしょう．そのため，「しょっぱい」だけを知っている人より「検潮儀」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します．では，「検潮儀」を知っていれば何語ぐらい知っていると仮定できるのでしょうか．その仮定の根拠となるのが，次に紹介する「単語親密度」です．

単語親密度データベース

「語のなじみ深さ」を評定実験によって数値化したものを「単語親密度」と呼びます．語に付与された数値が大きければ大きいほど，多くの人になじみのある語であり，数値が小さければ小さいほど，多くの人にとってなじみのない，あまり知られていない語であることを示します．

NTTでは20年以上前から単語親密度データベースなどの基盤的言語資源の構築に取り組んできました．過去に構築した約７万７千語からなる平成版の単語親密度データベースは，NTTデータベースシリーズ「日本語の語彙特性」として公開され，心理学や言語教育，言語聴覚療法分野などの基礎指標として幅広く活用されてきました．しかし，調査から時間が経ち，単語親密度自体が時代とともに変化した可能性もあり，新しく出現した語（「インターネット」や「コンビニ」など）に対応していないといった問題もありました．

そこで，再調査と新しい語の追加調査を実施し，令和版単語親密度データベースとして約16万３千語という過去最大のデータベースを構築しました[(1)]．さらに，平成版単語親密度からの変化を調査し，両者に強い相関があり，多くの語では20年以上経っても親密度に大きな変化がないことを確認しました．一方で，大きく変化した語も一部存在すること，どういった語が大きく変化したかを明らかにしました（**図１**）．

単語親密度を用いた語彙数推定方法

単語親密度を用いた語彙数推定では，親密度の高い語から低い語まで，何段階かの親密度の語をサンプリングして出題します（**図２**）．基本的に，知っていると回答された語

以上の親密度の語はすべて知っていると仮定して，語彙数を推定します．例えば，「しょっぱい」までなら約1500語，「検潮儀」まで知っていれば約13万９千語と仮定します．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にばらつきがあると考えられます．そこで，回答結果にロジスティック回帰曲線を当てはめ，知っている確率がちょうど50%になる語彙数を推定結果とします（**図３**）．

このようにすると，サンプリングした少数の語をチェックしてもらうだけで，回答者の語彙数を簡単に推定することができます．また，理論的には同じ親密度の語であれば何を出題しても良いため，出題する語の変更が容易で，さまざまなバリエーションのテストを作成することができます．もちろん，チェックしてもらう語は多いほど推定精度を向上させることができます．

本手法で推定できる語彙数の上限は単語親密度データベースのサイズに依存しますが，令和版単語親密度データベースの構築により，推定できる語彙数の上限が大きく上昇し，テストの汎用性を高めることができました．

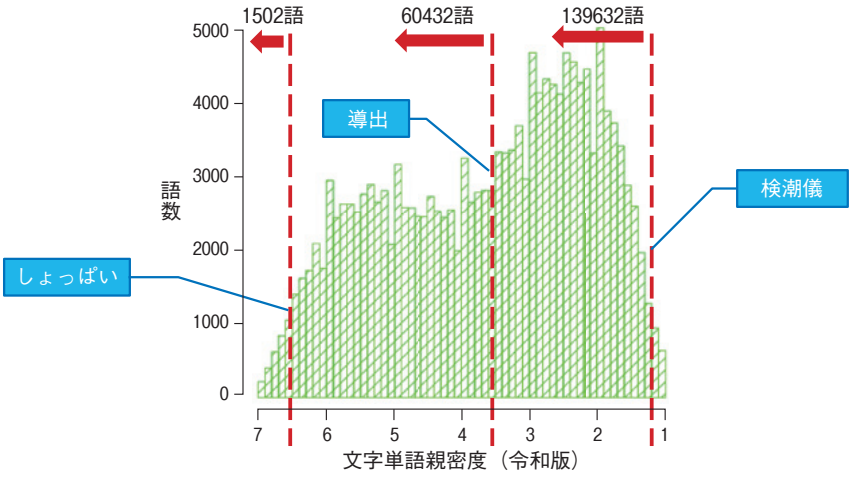

図２　単語親密度に基づくサンプリング

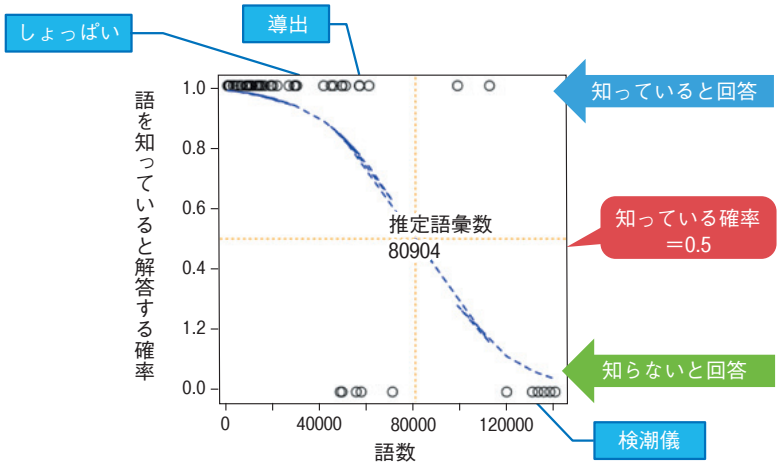

図３　語彙数推定方法

大規模な語彙数調査

これまでは児童・生徒を対象とした大規模な語彙数調査はほとんど実施されてきませんでしたが，私たちは，本手法を用いて，公立の小学生～高校生2800人以上を含む，約4600人の語彙数調査を行いました．

その結果，特に，小中学生では急激に語彙数が上昇すること，成人でも，年齢とともに語彙数が上昇することを確認しました（**図4**）．調査結果から，同じ学年でも生徒によって語彙数に大きなばらつきがみられるため，支援が必要な生徒を見つけるのに役立つと考えています[(2)]．

さらに，各学年・年齢における，単語親密度と語彙獲得状況（その語を知っている人の割合）との関係を分析しました（**図5**）．これは，各単語親密度の語を知っていると回答した人の割合（獲得割合）を各学年・年齢ごとに示したものです．どの学年・年齢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ります．一方，成人に比べて，小学生や中学生では，比較的親密度の高い語であっても，知っているかどうかの個人差が大きく，ばらつきがあります．こうした分析から，単語親密度を手掛かりとして，児童・生徒がこれから重点的に獲得するだろう語彙，あるいは獲得したほうが良い語彙を見つけていくことができると考えています．

また，学校での調査用とは別に，一般の方に試していただける令和版語彙数推定テストをオープンハウス2020に合わせて公開しました（http://www.kecl.ntt.co.jp/icl/lirg/resources/goitokusei/，2020年6月4日公開）．公開後約10日で3万人以上の方にお試しいただきました．よろしければ，皆様もお試しください．

テキストの難しさを測る

次に，テキスト側の難易度推定方法を紹介します．最初に取り組んだのは，絵本の難易度推定です．NTT研究所でこれまで取り組んできた幼児の語彙発達研究と組み合わせることで，語彙発達の解明や発達支援に寄与で

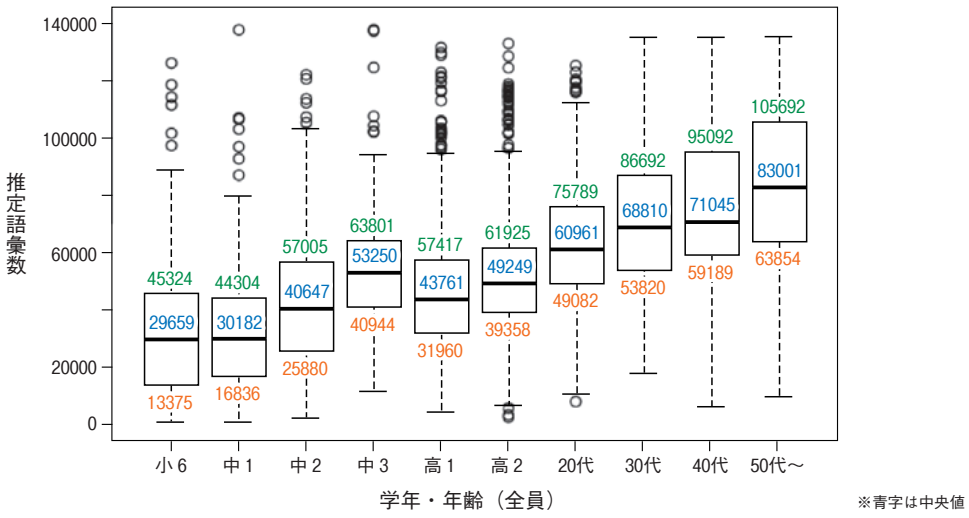

図4　学年・年齢の語彙数推定結果

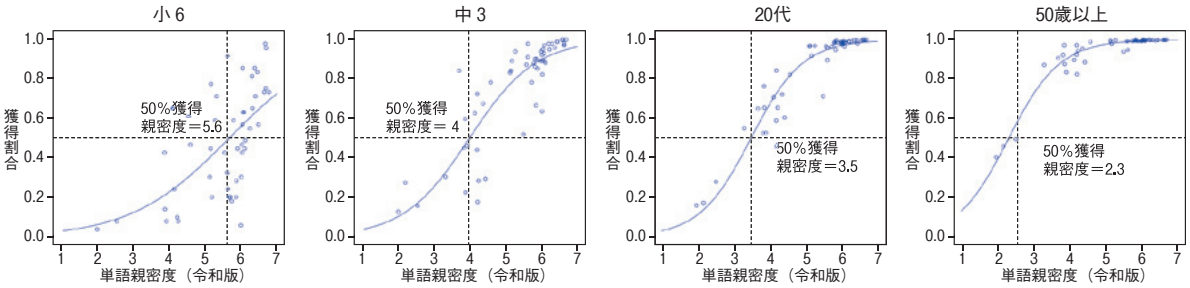

図5　単語親密度と語彙獲得状況の関係

きるのではないかと考えたからです．

テキストの難易度には，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します．絵本の場合，ひらがなを正しく解析することも必要です．例えば，「とうさん」が「父さん」か「倒産」かによって，難易度は大きく異なります．そうしたひらがな解析の精度向上に取り組み，幼児の語彙発達データベース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ることで，0～2歳，3歳，4歳，5歳という4クラス分類で87.8%という高精度な難易度推定を可能にしました[(3)]．また，一口に3歳向けといっても，2歳に近い3歳という場合もあれば，4歳に近い3歳という場合もあります．そのため，前後の年齢向けと推定してもあまり問題にはならないと考えられます．前後の年齢向けと推定しても良いとして評価すると，精度は96.7%となりました．つまり，提案手法は，対象年齢を大きく間違えることがほとんどなく，ロバストで信頼度の高い難易度推定が可能であるといえます．この難易度推定方法は，絵本検索システム「ぴたりえ」[(4)]で子どもにあった読みやすさの絵本を探すために使われています．

なお，テキストの難易度推定の研究では多くの場合，教科書の学年を推定できるかどうかで評価を行います．そこで，提案手法を適用して教科書の学年を推定する評価も実施したところ，小学1年生から中学3年生の9クラス分類で，98%以上という高い精度で学年を推定できました．つまり，絵本の難易度推定の精度向上のための工夫は，小学生以上のテキストの難易度推定にも有効であることが分かったのです．

NTT絵本・児童書コーパス

前述の難易度の推定には，電子データ化された絵本の文章のデータ（コーパス）を利用しています．しかし，絵本のコーパス自体もともとは存在しなかったため，その構築からスタートしました．実はこの部分がもっとも泥臭くて時間がかかる部分です．絵本では絵の中に文字が書かれることが多く，OCRによる文字認識もできません．結局，ほとんど人手で本文を入力することになりました．地道な作業の甲斐あって，NTTの絵本コーパスは，日本語の絵本6000冊以上，英語の絵本2500冊以上からなる世界に類をみない規模になりました．しかも今でも拡張中です．

さらに私たちは，構築したNTT絵本・児童書コーパスを用いて，絵本と幼児の語彙発達や感情発達との関係調査など，多くの新しい研究に取り組んでいます．

図 6　語彙力にあった英語コンテンツの推薦

今後の展開

今回紹介した人の語彙数やテキストの難易度の推定は独立に行っています．しかし，両者を組み合わせれば，定期的な語彙数の確認をしつつ，個人ごとに「ちょうど読めるくらい」や「少し頑張れば読めるくらい」のテキストを推薦することが可能になります．実際私たちは，英語の語彙数推定と難易度推定の研究も進めており，語彙力にあった英語絵本を推薦して学校の英語教育に活かす取り組みも始めています[(5)]（**図 6**）．

私たちは今後も，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，１人ひとりに合った育児・教育支援の実現をめざしていきます．

■参考文献

⑴ 藤田・小林：“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，2020.
⑵ 藤田・小林・山田・菅原・新井・新井：“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，2020.
⑶ 藤田・小林・南・杉山：“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.
⑷ 藤田・服部・小林・奥村・青山：“絵本検索システム「ぴたりえ」～子どもにぴったりの絵本を見つけます～，”自然言語処理，Vol. 24, No. 1, pp. 49-73. 2017.
⑸ 藤田・服部・小林・納谷：“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，2020.

藤田　早苗

自分の当たり前と，他人の当たり前は違います．忘れがちなことですが，自分や大人の「当たり前」を押し付けるのではなく，１人ひとりに合った育児・教育支援の手助けとなるよう今後も取り組んでいきます（自戒を込めて）．

◆問い合わせ先
NTTコミュニケーション科学基礎研究所
協創情報研究部
言語知能研究グループ
TEL　0774-93-5020
FAX　0774-93-5026
E-mail　cs-liaison-ml＠hco.ntt.co.jp