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～

Which word is more difficult for you, “car” or “vehicle” ?

— Estimation of text readability and human vocabulary size —

協創情報研究部

藤田　早苗 Sanae Fujita

プロフィール

NTT コミュニケーション科学基礎研究所 協創情報研究部 主任研究員。1999年奈良先端科学技術大学院大学 情報科学研究科 修士課程修了。同年、NTTに入社。博士（工学）。自然言語処理研究に従事。言語処理学会、情報処理学会各会員。3人の子どもの成長とともに、絵本の難易度推定、学齢期の語彙数調査、英語学習支援と研究テーマを広げています。

文字を覚えたての子どもが自分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1の時とても苦労して読んだ英文が、大学生になる頃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、読み手の知識を無理なく増やしていけるかもしれません。しかし、「ちょうど良い難しさ」を判断するのは簡単ではありません。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るからです。

人の語彙数調査と推定方法

人側の知識の一つとして語彙力があります。NTTでは20年以上前から、人の語彙数の調査や推定に取り組んできました。

幼児を対象とする場合、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。提示する語は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、何語知っていると推定するかがポイントとなります。例えば「銀行」と「地歩」だと、「地歩」の方が知っている人は少ないでしょう。そのため、「銀行」だけを知っている人より「地歩」も知っている人の方が語彙数は多いと思われます。では「地歩」も知っていれば、「何語知っている」と推定できるのでしょうか？

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」です。NTTでは20年以上前に約7万7千語の単語親密度を調査、それを元に公開した語彙数推定テストは多くの方に利用されてきました。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、16万3千語というより大規模な単語親密度データベースを再構築しました[2]。

データベースを元に令和版語彙数推定テストを作成、4千人以上を対象に調査を行ったところ、小学6年生で約2万語、大人で約5万語の語彙数があることがわかりました（図1）。また、同じ学年でも生徒によって語彙数に大きなばらつきがあることから、支援が必要な生徒を見つけることにも役立つと考えています[3]。

テキストの難易度推定

テキスト側の難易度については、まず絵本の難易度推定から取り組みました。本研究所の主要テーマでもある幼児の語彙発達の解明、発達支援などにも寄与できると考えたからです。しかし、絵本には電子データ化されたデータベース（コーパス）自体が存在せず、その構築からスタートしました。人手でほとんどの本文テキストを入力するという地道な作業の結果、NTTの絵本コーパスは、日本語6000冊・英語2000冊を超える、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。

さて、いよいよ難易度推定です。テキストの難易度には、語彙の難しさ、文構造の難しさの両方が影響しますので、それぞれを適切に反映する特徴量を見出しました。加えて、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。こうした特徴量選択やひらがな解析の工夫により、高精度な難易度推定を可能にしました[1]。また、この精度向上の工夫は、教科書など他分野のテキストの難易度推定にも有効であることが分かってきました。

今後の展開

今回ご紹介した「人の語彙数推定」と「テキストの難易度推定」を組み合わせ、定期的な語彙数の確認をすれば、一人ひとりがその時、「ちょうど読める・少し頑張れば読める」テキストを推薦することが可能になります。実際我々は、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（図2, [4]）。

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・教育支援の実現をめざしていきます。

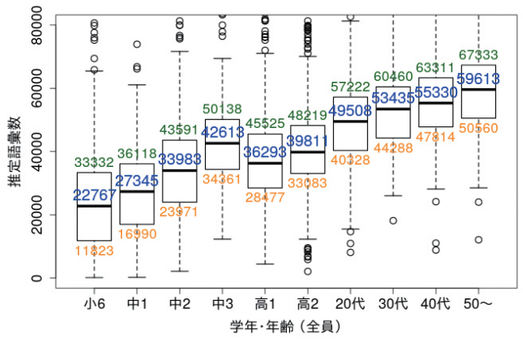

図1　小学6年生から成人を対象とした語彙数推定結果（約4600名）

図2　英語の語彙数推定画面

▼関連文献

[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.