クラウドソーシングを用いた習得時期の想起質問に基づく単語難易度データベースの構築

水谷 勇介　　河原 大輔　　黒橋 禎夫
京都大学 大学院情報学研究科
{mizutani, dk, kuro}@nlp.ist.i.kyoto-u.ac.jp

1　はじめに

単語に難易度を付与することは、自然言語処理や言語の学習において大きな意義がある。例えば、単語の難易度を踏まえた学習や教材の作成 [1,2]、テキスト平易化のための単語の言い換え [3, 4] 等に貢献する。しかし、国語教育では漢字の難易度の基準 (配当学年) は定められているが、単語の難易度に関する基準は定められていない。例として漢字問題の作問を考えると、「委細」という単語を構成する「委」と「細」はそれぞれ小学校三年と二年で習う漢字であるが、「委細」の意味を理解している小学三年生は少ないと思われるため、「委細」を三年の問題にするのはふさわしくない。このように、単語に難易度が設定されていないため、作問者の内省をもとに、学年やレベルに合った単語を選別して問題を作成しているのが現状である。

本研究では単語難易度のデータベースを構築する。単語難易度は、**単語を理解し、使い始めた時期** (以下、習得時期) と定義する。習得時期を求めるためには、小学生の読書感想文のように学年が区別できるテキストを用いて単語頻度を計数する方法が考えられるが、テキスト収集のコストの観点から実際に行うことは難しい。本研究ではクラウドソーシングを用いて習得時期の収集を行う。クラウドワーカーの大多数は大人であり、正確な習得時期は覚えていないと考えられるが、およその習得時期を想起させる質問をすることによって習得時期を得る。収集した習得時期データと、単語難易度と関係があると思われる既存のリソースとの比較を行い、習得時期データの妥当性を検証する。

2　クラウドソーシングによる単語習得時期の収集

クラウドソーシングによる単語習得時期の収集は、図 1 に示すように、クラウドワーカーに習得時期を想起させる質問をすることによって行う。回答は 5 択の選択式とし、選択肢としては「小学校になる前」「小学校低学年」「小学校高学年」「中学生になった後」「単語の意味を知らない/見聞きしたことがない」を設定する。なお、本研究で扱う単語は、形態素解析器 JUMAN++ の単語辞書に掲載されている単語とし、約 26,000 語を対象とする。日本語 Web98 億文の形態素解析済みデータから各単語の頻出表記を求め、クラウドソーシングで質問する際に使用する表記とする。

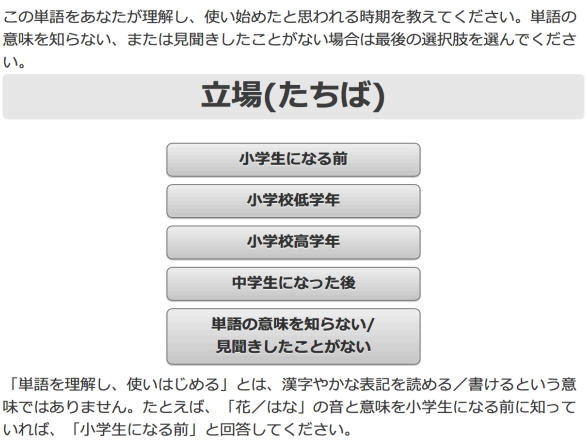


図 1: 習得時期の想起質問と回答の選択肢

クラウドソーシングは 5,000 単語ずつに分けて行い、合計 6 回実施した。1 タスク (ワーカーが一度に回答する問題数) は 20 問からなり、1 回のクラウドソーシングごとに 1 人最大 10 タスクを行うことが可能である。各単語の習得時期は 20 人から収集した。1 つのタスクの回答が全て同じであるような質の悪いワーカーを除外した結果、1 回のクラウドソーシングにつき約 500 人のワーカーが参加し、6 回のクラウドソーシングで延べ 3,121 人が参加した。

「小学生以前」から「単語の意味を知らない/見聞きしたことがない」まで順番に 1〜5 の値を与え、各単語について 20 人の回答の平均値を求めた。その結果

習得時期	語数	単語例
1.0-1.5	196	赤い、おでこ、友達、上がる、チョコ、小学校、月曜、怖い、泣き声、クリスマス、女の子、こんにゃく、投げる、ハンカチ、コロッケ、ウンチ、増える、帽子、ピアノ、同じだ、暑い
1.5-2.0	1,771	ゴム、ピーナッツ、三輪車、ボート、パンク、殴る、来年、水漏れ、キラリ、人助け、悔しい、ダンス、次に、デザート、探し物、夕立、ラジコン、片方、隣、でも、手本、元日、回転
2.0-2.5	4,586	へその緒、右腕、磁石、静まる、毎度、白黒、長方形、クラブ、そそっかしい、睨む、人口、雑だ、引き継ぐ、昇れる、発熱、ウ冠、一人前、突き刺さる、常に、遠吠え、音程、見方
2.5-3.0	8,090	無鉄砲だ、卵白、物事、師弟、切り上げる、本店、泣きはらす、ハネムーン、焦げ付ける、連想、茶漉し、視覚、さらし粉、通勤、垂線、原則、スキーヤー、使用人、包囲、代謝
3.0-3.5	8,817	そそくさ、臨終、機器、不景気だ、風習、症例、明滅、手負い、経由、浮気だ、両党、紀行、凸面鏡、報ずる、増員、兆せる、割賦、門徒、臥せる、闊歩、画引き、失脚、原簿、実装
3.5-4.0	3,100	背任、縦走、特使、克明だ、国務、助成、芍薬、集成材、増収、滋養、追徴、提訴、訓辞、惰眠、享受、韻文、貸付、バント、小兵、新前、騙れる、御息所、輻射、肋木、蛮行、編纂
4.0-4.5	173	僧籍、公判廷、招請、消長、運筆、草葺、居竦まれる、ゲノム、烈震、巨視的だ、随意だ、鳥瞰図、料簡、彼我、同床異夢、詠嘆、春本、一事不再理、開削、極光、篆刻、頁岩
4.5-5.0	3	山家、冊封、枝折戸

表 1: 習得時期ごとの単語数と単語例 (単語の並びは習得時期順になっている)

を表 1 に示す。表からわかるように、単語難易度として比較的妥当な結果が得られたと考えている。「この単語をあなたが理解し、使い始めたと思われる時期」という表現で質問したことによって、当然その時期を覚えていないが、想起して答えた時期が我々の単語難易度の感覚と相関があったと考えられる。以下では、収集したデータを**習得時期データ**、各単語に対して得られた習得時期の値を単に習得時期と呼ぶ。

3 習得時期データと既存リソースとの比較

習得時期データの単語難易度としての妥当性を検証するために、単語難易度と関係があると思われる既存のリソースとの比較を行う。それらのリソースは大きく分けて次の三つに分類することができる。

1. 教育機関が作成したリソース
2. コーパスから自動推定したリソース
3. アンケート調査により作成したリソース

以下の節ではそれぞれのリソースとの比較について述べる。なお、習得時期データと各リソースとのピアソン相関係数を表 2 に示す。

節	リソース	相関係数
3.1	教科書カテゴリ	0.543
	漢検難易度	0.598
	日本語能力試験の級	0.565
3.2	$\log_{10}$(単語頻度)	0.517
3.3	NTT 文字音声単語親密度	0.716
	NTT 文字単語親密度	0.706
	NTT 音声単語親密度	0.647

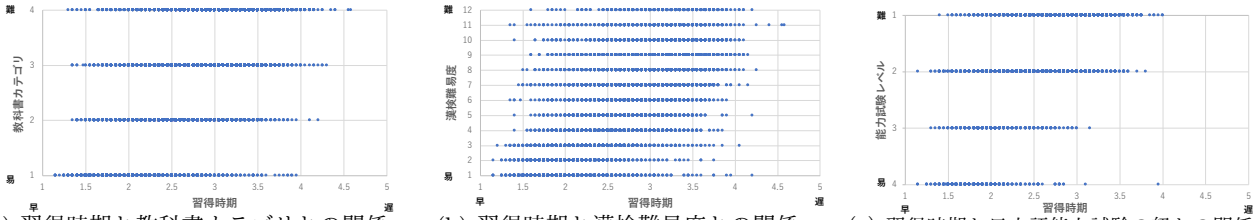
表 2: 習得時期データと各リソースとのピアソン相関係数

3.1 教育機関が作成したリソース

教育機関が作成したリソースとして、教科書、日本漢字能力検定 (漢検) の過去問、日本語能力試験の出題基準語彙の 3 つを対象に習得時期データとの比較を行う。なお、教科書と漢検の過去問は、各漢字についての配当学年や配当級の設定はあるが、各単語についての難易度のレベル設定は存在していない。日本語能力試験の出題基準語彙は、単語について難易度を設定したものであるが、日本語を母語としない人を対象としたものである。

3.1.1 教科書

教科書から抽出した難易度付きのリソースとして、教科書コーパス語彙表を用いる。教科書コーパス語彙表は、教科書コーパスから抽出した語彙リストであり、小学校前半、小学校後半、中学校、高校の 4 つのカテゴリにおける単語出現頻度が掲載されている。この語彙表から、単語ごとに出現した最も低い学年のカテゴリを抽出した。これを教科書カテゴリと呼ぶ。習得時期データと共通する約 14,000 語について、習得時期と教科書カテゴリの関係を図 2(a) に示す。図では、小学校前半から高校までに 1 から 4 の値を与えている。教科書に使用できる単語に限りがあるため、低学年の教科書に使用できる単語がたまたま高学年の教科書でのみ使用されることがある。図 2(a) の左上の領域に分布している教科書カテゴリの値は高いが習得時期が早い単語がこれに該当しており、習得時期の方が単語難易度として妥当であると考えられる。



(a) 習得時期と教科書カテゴリとの関係　(b) 習得時期と漢検難易度との関係　(c) 習得時期と日本語能力試験の級との関係

図 2: 習得時期と教育機関が作成したリソースとの関係

3.1.2 漢検の過去問

漢検は、漢字能力を測定する技能検定である。10 級～1 級の 12 段階の級からなり、10 級が小学一年生修了程度で、1 級がもっとも難しい。本研究では、漢検の問題文中に使用された単語について、出現した最も簡単な級を抽出した。これを漢検難易度と呼ぶ。平成 4 年から 28 年までに出題された漢検の問題文のうち、読み問題と書き問題 (選択式の問題は除外) に出現する単語を抽出した。習得時期データと共通する約 12,000 単語について、習得時期と漢検難易度の関係を図 2(b) に示す。教科書と同様に、出題に使用できる単語に限りがあるため、低い難易度の単語がたまたま高い難易度の問題文でのみ使用されていることがある。左上の領域に分布している漢検難易度は高いが習得時期が早い単語がこれに該当している。例えば、「怪獣」や「しゃがむ」という単語は 1 級の問題文でのみ使用されているため漢検難易度は最も難しい 12 となる。一方、習得時期はそれぞれ 1.85 と 1.6 (小学校低学年よりすこし下) であり、単語難易度の指標として妥当な値であると言える。また、難易度の低い級ではひらがな表記の単語が多く、単語分割を行う際に同音異義語を誤って認識してしまうことがある。例えば、「水槽 (すいそう)」という単語が平仮名で表記されているために、「吹奏」という単語として抽出されていた。右下の領域に分布している漢検難易度は低いが習得時期が遅い単語がこれに該当している。

3.1.3 日本語能力試験の出題基準語彙

日本語能力試験は、日本語を母語としない人の日本語能力を測定し認定するための試験である。2010 年から実施されている (新) 日本語能力試験に用いられている出題基準は非公開のため、2009 年まで用いられていた旧版の出題基準語彙に関するデータを比較のためのリソースとして用いる。旧版は 1～4 級まであり、4 級が最も簡単で、1 級が最も難しい。このデータには、各級の出題基準語彙が記載されている。習得時期データと共通する約 5,000 単語について、習得時期と日本語能力試験の級との関係を図 2(c) に示す。図

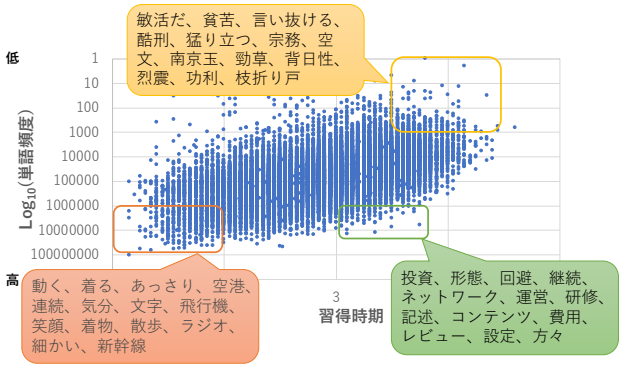


図 3: 習得時期と単語頻度との関係

2(c) の左上の領域には、「裸足」「抱っこ」「なぞなぞ」などのように、級は高いが習得時期が早い単語が存在している。これらは、日本人には簡単でも、日本語を母語としない人には難しいと思われる単語である。

3.2 コーパスから自動推定したリソース: 単語頻度

コーパスから単語の出現頻度を計数し、その値を単語難易度として用いるということが行われてきた [5]。本研究では、日本語 98 億文からなるウェブコーパスの形態素解析済みデータを用いて単語頻度を計数し、それを習得時期データと比較する。習得時期データと共通する約 25,000 単語について、習得時期と単語頻度との関係を図 3 に示し、いくつかの領域についてはどのような単語が実際に分布しているかを例示する。ただし、単語頻度は対数をとっている。習得時期が遅くてもよく使う単語などが存在するため、習得時期と単語頻度の相関は高くないことがわかる。例えば、「投資」や「研修」は社会で頻繁に使われ、ウェブコーパス上での出現頻度からも高頻度な単語といえるが、習得時期はそれぞれ 3.15 と 3.05(小学校高学年よりすこし上) である。この結果は、先行研究の報告 [5] と合致している。

3.3 アンケート調査により作成したリソース: NTT 単語親密度

アンケート調査によって作成された単語難易度に関係するリソースとして、「NTT データベースシリーズ

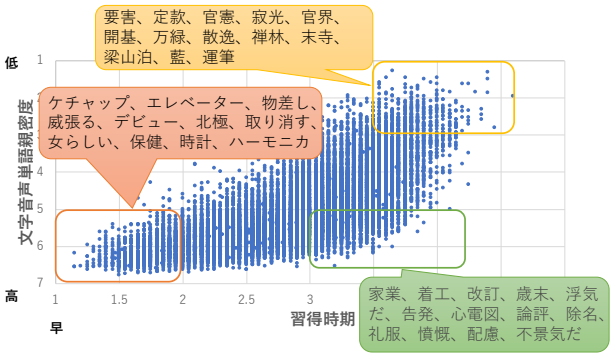


図 4: 習得時期と NTT 文字音声単語親密度との関係

日本語語彙特性 単語親密度 [6]」(以下、NTT 単語親密度) がある。これは、単語親密度を単語のなじみの程度と捉え、それをアンケート調査で収集したものである。被験者数は 18 歳以上 30 歳未満の 40 人から、音声のみと文字のみ、音声と文字の両方を刺激として使用した 3 つの実験手法がある。主に『新明解国語辞典』第四版の見出し語を使用しており、88,569 単語の親密度を 7 段階尺度 (1:低 → 7:高) で収集している。単語には「ぶどう」と「葡萄」のように複数の表記方法を持つものがあるが、ここでは頻出表記に該当する単語について議論を行う。

3 つの実験手法のうち、習得時期データと最も相関があったのは文字音声親密度なので、以降の比較では文字音声親密度を用いる。習得時期データと共通する約 21,000 単語について、習得時期と文字音声親密度の関係を図 4 に示し、いくつかの領域についてはどのような単語が実際に分布しているかを例示する。図 4 の右下の領域に、単語親密度は高いが習得時期が遅い単語が分布している。例えば「不景気だ」の単語親密度は 6.125、「浮気だ」は 6.062 であり、これらは世の中で頻繁に見かける単語であるため、なじみがあるが、単語難易度は中程度以上と思われる。それぞれの単語の習得時期は 3.15 と 3.25 (小学校高学年よりすこし上) であり、単語難易度の指標として妥当な値であると言える。

3.4 総評

表 2 によると、習得時期データともっとも高い相関係数をもつのは NTT 単語親密度である。双方ともアンケート調査で収集しており、単語難易度に直結する習得時期やなじみに対する人々の感覚をうまく収集できていると考えられる。しかし、NTT 単語親密度は、難しいと感じられる一部の単語に対して親密度が高くなっており、単語難易度の指標として用いるには問題がある。

NTT 単語親密度に次いで相関係数が高かったのは、漢検難易度など、教育機関が作成したリソースである。これらのリソースは、専門家が学習者のレベルに応じて単語や文章を選定しているため、良質のリソースであると考えられる。しかし、教科書や問題に使用できる単語数に制限があるため、易しいと感じられる一部の単語に対して高い難易度が設定されてしまう問題がある。

習得時期は、これらの単語に対しても適切な値が獲得されており、単語難易度の指標として妥当なものであると考えられる。

4 おわりに

本研究では、クラウドソーシングを用いて単語の習得時期に関する想起質問を行い、単語難易度データベースを構築した。得られた各単語の習得時期は、我々が考える単語の難易度と密接に関係していることがわかった。今後は、この習得時期を用いた漢字の学習アプリを開発し、日本語学習を支援する研究に役立てていきたい。また、本研究で構築した単語難易度データベースは公開予定である。

謝辞

本研究は京都大学と（公財）日本漢字能力検定協会の研究プロジェクト「人工知能（ＡＩ）による漢字・日本語学習研究」のもとで実施された。（公財）日本漢字能力検定協会からの研究助成に感謝致します。

参考文献

[1] 藏培慶, 小林伸行, 椎名広光. 単語難易度推定による中日単語学習システム. 言語処理学会第 20 回年次大会発表論文集, pp. 113–116, 2014.

[2] 中西聖明, 木藤善信, 木村祐介, 椎名広光, 北川文夫. 日本語の単語難易度推定による VOD 講義の難易度推定. Technical report, Information Processing Society of Japan, 2011.

[3] 梶原智之, 小町守. Simple PPDB: Japanese. 言語処理学会第 23 回年次大会発表論文集, pp. 529–532, 2017.

[4] 田中英輝, 熊野正, 後藤功雄, 美野秀弥. やさしい日本語ニュースの制作支援システム. 自然言語処理, Vol. 25, No. 1, pp. 81–117, 2018.

[5] 梶原智之, 山本和英. 高頻度語は平易語なのか？, 2014.

[6] 天野成昭, 近藤公久. 日本語の語彙特性, 第 1 巻. 三省堂, 2000.