

日本語学習支援を目的とした漫画の台詞の難度の判定

Recommendation of Spoken Lines of Comic
by Evaluating the Degree of Difficulty for Learners of Japanese

西原陽子*1
Yoko Nishihara

SHAN, Junjie*2
Junjie Shan

山西良典*1
Ryosuke Yamanishi

福本淳一*1
Junichi Fukumoto

*1 立命館大学情報理工学部

College of Information Science and Engineering, Ritsumeikan University

*2 立命館大学大学院情報理工学研究科

Graduate School of Information Science and Engineering, Ritsumeikan University

The purpose of this research is to propose an evaluation method of the degree of difficulty of spoken lines of comic of Japanese language learners. Many foreigners who are not Japanese natives often enjoy Japanese comics and cartoons and learn Japanese. In comics and cartoons, conversations of spoken lines by characters are included. So, the learners can train them to make conversations in Japanese by reading/listening the spoken lines. Some spoken lines are easy to understand for learners even if they are beginners of Japanese. However, there may be spoken lines that are hard to understand for the beginners because they do not have enough knowledge of Japanese. Therefore, we propose the method to support Japanese language learners by providing appropriate spoken lines for learners' level.

1. はじめに

日本の漫画やアニメは、日本国内だけでなく、海外でも頻繁に楽しられている。漫画であれば、その国の言語に翻訳したものが海外で出版されており、アニメであれば、その国の字幕がつけられたものが提供されている。翻訳や字幕を通じて、日本の漫画やアニメが楽しられている。

日本語の勉強をする人の中には、漫画やアニメを楽しむ中で勉強をしている人もいる。漫画やアニメは、キャラクター達の会話や、心の中で考えたことを出す独白が、頻繁に出てくる。もちろん日本語を勉強するための教科書もあるが、会話や独白の例がふんだんにあるわけではない。漫画やアニメでは、会話や独白といった、話し言葉を十分に学ぶことができる。

漫画やアニメのセリフは、対象年齢以上の日本人ならば理解できるように構成されているが、外国人にとっては理解が簡単なものと、難しいものが存在する可能性がある。例えば、朝の通学時のシーンで「おはよう」「おはようございます」と学生と先生が挨拶をし合うセリフは、日本語の初學者でも理解できる可能性が高い。しかし、「その使用人ってのが あいつよでも 3日間問いただしても 知らぬ存ぜぬの一点張り...」(ベルモンド Le Visiteur p.18, (c) 石岡ショウエイ)と「知らぬ存ぜぬ」「一点張り」などの表現は、ある程度学習を済ませた人でないと理解できない可能性が高い。日本語の全くの初學者と、ある程度学習を済ませた人では、内容を理解できる台詞と、そうでない台詞があると考えられる。台詞の難度が評価され、学習者のレベルに応じた台詞が与えられることが望ましい。

さらに、話し言葉は、同じフレーズであっても、使われる状況によって、表現が変化する。この例として「失礼します」というフレーズがある。部屋に入る際に、その部屋の中にいる人によって、表現を変化させる必要がある。生徒が部屋に入る場合、中にいるのが馴染みの先生であれば「失礼します」と語尾を伸ばしても良いが、馴染みのない先生であれば「失礼します」と語尾を伸ばすのは避けた方が無難である。敬意を表したければ「失礼いたします」と謙讓表現にすべきである。また、

話し手が小さい子であれば「失礼しまちゅ」も許容できるが、大人の大人が使う事はあり得ない。同じフレーズでも、カジュアルな表現と、フォーマルな表現がある。これらは状況によって、使う表現を変化させることが望ましい。

そして、漫画やアニメは、それぞれ扱っている話の内容によってジャンル分けがされている。ジャンルとしては、日常系、推理もの、スポーツもの、冒険ものなどがある。同じ話し言葉であっても、実際の日常生活でよく使われる物と、使われない物があると考えられる。例えば、日常系漫画であれば、日常生活を題材としているため、日常生活でも使われる言葉が多いと考えられる。一方で、冒険漫画であれば、創作の世界の中をキャラクター達が冒険していくストーリーを描いているため、キャラクターの技の名前や、魔法の名前など、日常生活では使われない言葉が多いと考えられる。日常生活で使われる単語とそうでない単語を分けることが望ましい。

すなわち、漫画を用いて話し言葉を学ぶ上では3つの問題点があると考えられる。3つの問題点とは、

問題点1 台詞の難度が評価されていない。

問題点2 同じフレーズでも、カジュアルな表現とフォーマルな表現があり、使い分けを学ぶことが難しい。

問題点3 日常生活で使う言葉と、使われない言葉が分けられていない。

である。

そこで、本研究では、日本語の学習者が話し言葉を学ぶために漫画を用いることを想定し、学習者に適した漫画を推薦することを目標とする。本研究では、一例として、中国人の日本語学習者を対象として、研究を進める。

1.1 日本語の学習と能力評価

中国人の日本語の学習者の多くは、将来的に日本に留学するために、日本語を勉強していることが多い。国費留学生や私費留学生として、大学に入学する際、大学からは一定のレベルの日本語能力を有することを求められることが多い。こ

の能力の評価に使われている試験として、日本語能力試験 (<http://www.jlpt.jp/index.html>) がある。日本語能力試験では、日本語を学ぶ外国人を対象として、他者とコミュニケーションを取る上での課題を遂行できるかどうかを評価する。評価点は2つに分かれ、(1) 日本語の文字、語彙、文法など日本語の知識を評価する点と、(2) 日本語の知識を実際コミュニケーションでどの程度使えるかを評価する点がある。試験は、(1) の評価のために言語知識を問う問題、(2) の評価のために読解問題(文章を読んで、適切な選択肢を選ぶ)、聴解問題(音声を聞いて、適切な選択肢を選ぶ)の3つに分けられている。試験はN1からN5まで、5段階に分けられていて、N1が最も難しく、N5が最も簡単となっている。^{*1} 話し言葉の学習に漫画を用いることにより、(1) の言語知識を問う問題や、(2) の読解問題の勉強ができると考えられる。

この試験用の勉強をするための教科書がある。多くの中国人が利用している教科書として、日本語能力試験語法問題対策 [日本語能力試験語法問題対策] や、中日交流標準日本語 [中日交流標準日本語初級] がある。中日交流標準日本語は、日本人が用いる英語の教科書と良く似ており、学ぶ文法ごとにレッスンが分けられており、各レッスンの冒頭には会話の例が示されている。文法の説明がなされ、その後、新しく出てきた単語や、関連する単語の紹介があり、最後に文法や単語の練習問題がある。レッスンは基礎的な内容から始まり、徐々に応用的な内容へと進んで行く。また、レッスン番号は試験範囲と対応している。難度の評価にこれらの教科書を用いる事で、試験出題範囲に対応して、漫画の推薦が可能になると考えられる。

1.2 問題点1から3の解決方法

単語や文法の難度を評価するために、単語や文法が試験範囲となっている、日本語能力試験の試験レベルN1からN5の数値を用いる。これにより、問題点1の解決に取り組む。テキストの難度を評価する研究は、英語や日本語で書かれたテキストの難度を評価する研究がある [高木 91]。近藤らの研究では、教科書コーパスを用いて、文字 unigram の言語モデルを作成し、テキストを難易度別に分類する手法を提案している [近藤 08]。本研究では、難度判定の対象が漫画のセリフである。近藤らの手法ではある程度の長さを持ったテキストを対象としている。より短いテキストの難度判定に本研究では取り組む。

カジュアルな表現とフォーマルな表現の使い分けを学ぶために、同じフレーズが使われている台詞をクラスタリングして、提供する。これにより、問題点2の解決に取り組む。ビジネスメールにおいて、カジュアルな表現とフォーマルな表現は使い分けが求められている [山崎 13]。英語では、用いる文型や表現によりカジュアルとフォーマルの使い分けがされている。本研究では、日本語のカジュアルな表現とフォーマルな表現の分類に取り組む。

本研究では、日常生活で使う話し言葉のサンプルとして、中日交流標準日本語などの中国人が日本語の勉強に用いている教科書を利用する。これにより、問題点3の解決に取り組む。

本論文では初めに、難度の評価を行う手法を提案する。

2. 提案手法

提案する手法では、漫画の台詞のテキストデータを入力とする。漫画1冊分のセリフデータを入力し、1つの吹き出しに入っているセリフをひとまとまりとして、区切る。区切られた

セリフの難度を評価し、1冊の漫画全体として、どの難度のセリフが入っているかを出力する。

2.1 入力：漫画の台詞のテキストデータ

漫画の台詞を書き起こしたテキストデータを入力とする。漫画では、吹き出し中にキャラクターのセリフが記載されていることが多い。一つの吹き出し中のセリフをひとまとまりのセリフデータとする。

本研究では Manga109 [Matsui15]

(<http://www.manga109.org/index.php>) で提供されている漫画セリフデータを用いる。「ラブひな 1巻 (c) 赤松 健」、「ベルモンド Le Visiteur 1巻 地下に潜む男 (c) 石岡 ショウエイ」、「プレイヤーは眠れない (c) 黒沢 哲哉, 正木 秀尚」、「はるかりフレイム (c) 伊藤 伸平」と4種類の漫画セリフデータのファイルがある。ファイル内は、漫画のページごとに行が分かれており、1列目にページ番号、2列目にセリフのテキストがある。テキストは、主にセリフのまとまりごとに改行されている。改行を区切りとして、セリフデータを小さく分け、ひとまとまりのセリフデータとする。

2.2 セリフの難度の判定

入力された漫画のセリフデータのテキストを用いて、セリフの難度を判定する。セリフはひとまとまりごとに、難度を判定する。一般に、一つのセリフの中に、難しい文法が多数含まれ、また難しい単語が多数含まれるならば、そのセリフを理解するには多くの事前知識が必要となる。したがって、このようなセリフは難度が高いと評価される。一方、一つのセリフの中に、簡単な文法が数個だけ含まれ、また簡単な単語が数個だけ含まれるならば、そのセリフを理解するには事前知識は多くはいらない。したがって、このようなセリフは難度が低いと評価される。したがって、本研究では、難度を、使われている文法と単語の難しさ、およびその個数により評価する。セリフ s_i の難度 d_i は、使われている文法 N1 から N5 の個数 $g_{i1} \dots g_{i5}$ 、と使われている単語 N1 から N5 の個数 $w_{i1} \dots w_{i5}$ を用いて、式 (1) により与えられる。

$$d_i = \sum_{j=1}^5 (6-j) \times g_{ij} + \sum_{j=1}^5 (6-j) \times w_{ij} \quad (1)$$

式 (1) の一つ目の項は、文法の難度の和、二つ目の項は、単語の難度の和を表す。式 (1) の値が高い方が、難度が高いと評価される。難度は、個々の文法と単語の難度の総和になると考えられるため、平均はとらない。

文法や単語の難しさを決めるために、日本語の教科書である日本語能力試験語法問題対策の1級、2級、3級を用いる。1級がN1の試験、2級がN2の試験、3級がN3の試験となる。教科書では、試験範囲ごとに文法と単語が列挙されているため、それを書き起こした文法のテキストと、単語のテキストを用意する。テキスト中で文法は正規表現を用いて記載し、単語はそのまま記載する。N1、N2、N3の文法の正規表現の一例を表1に示す。用意した文法のテキストでは、N1には103件、N2には180件、N3には114件の文法の正規表現がある。

3. まとめと今後の課題

本論文では、日本語の学習者が話し言葉を学ぶために漫画を用いることを想定し、学習者に適した漫画を推薦するための手法の提案を目標とした。日本語の学習者の一例として、中国人の学習者を対象とした。漫画を用いて話し言葉を学ぶ上での

*1 著者が所属する大学では、外国人留学生にはN1の能力を要求している。

問題点として、(1) 台詞の難度が評価されていない、(2) 同じフレーズでも、カジュアルな表現とフォーマルな表現があり、使い分けを学ぶことが難しい、(3) 日常生活で使う言葉と、使われない言葉が分けられていない、と3つの点を示した。本論文では初めに(1)の解決に取り組む事とした。提案した手法は、漫画のセリフデータを入力とし、セリフのひとまとまりごとに難度を評価し、出力するものである。今後、手法の評価を行い、残り2つの問題点を解決し、漫画を日本語学習に用いる学習者に使われるアプリケーションとして公開していく。

謝辞

本研究では、東京大学 大学院情報理工学系研究科 電子情報学専攻 相澤・山崎研究室により取りまとめられた Manga109 の漫画セリフデータを利用しました。ここに記して謝意を申し上げます。

参考文献

- [日本語能力考試語法問題対策] 重野美枝他, 完全掌握 3 級日本語能力考試語法問題対策, 外語教学与研究出版社, (2009).
- [中日交流標準日本語初級] 中日交流標準日本語初級, 中国人民教育出版社, (2006) .
- [Matsui15] Yusuke Matsui, Kota Ito, Yuji Aramaki, Toshiko Yamasaki, and Kiyoharu Aizawa, Sketch-based Manga Retrieval using Manga109 Dataset, arXiv (2015)
- [山崎 13] 山崎克之, 電子メールの英語とマナー, 電子情報通信学会 通信ソサイエティマガジン, Vol. 7, No. 1, pp.54-58 (2013).
- [高木 91] 高木裕子: 速読用読解教材開発に向けて -リーダビリティ研究を基礎にして, 関西外国語大学留学生別科目日本語教育論集, Vol.1, pp. 66?85 (1991).
- [近藤 08] 近藤陽介, 松吉俊, 佐藤理史, 教科書コーパスを用いた日本語テキストの難易度推定, 第 14 回言語処理学会年次大会, pp.1113-1116 (2008).

表 1: N1 と N2 の範囲の文法の正規表現

レベル	文法の正規表現
N1	.+の?いかん [いなの](限 かぎ)り [だで] がてら .+かたがた ずには [済す] まな (い かった) で?すら ただ.+のみ っ(放—ばな)し (.+であれ)+ (では じゃ) あるまいし
N2	として [はもの]? と (共 とも) に に (応 おう) じ [てた]? に [代か] わ (って り) [をは][問と] わず を (始 初 はじ) め (とする)? を (元 もと) に (して)? お (陰 かげ)[でだ]? [代替か] わりに .+[くく] らい (だ です)?
N3	(後 あと) で (頂 いただ)(く き いて いた) (貰 もら)(う い って った) よ?うと (思 おも)(う い って) よ?うと (する します して した) お?.*?(下 くだ) さ (る い って った) お?.*?(御座 ござ)(る います) [おご]?.*?(する します) が (出来 でき)(る ます ない なかった ません)