N-025

外国人向け日本語学習支援システムのための形態素解析
Morphological Analysis for Foreigner's Japanese Learning System

森山 柊平† 絹川 博之†
Shuhei Moriyama Hiroshi Kinukawa

1. はじめに

外国人を対象とした日本語学習においてコンピュータが多く利用されるようになってきた．しかし，外国人日本語学習者が作成した文章を添削するシステムはほとんど見られず，日本語教師等の人手によって添削されているのが現状である．そのため，学習者が独学で文章作成を学習することは困難である．初級日本語学習者は，単語，漢字の読み書き，文法，文章作成など，様々な分野の学習をする必要がある．その中でも文章作成は，その添削に関してコンピュータで支援する必要があると考えられる． そこで我々は外国人学習者が独学で文章作成を学習できることを目標として日本語学習支援システムを開発している．現段階では，対象を初級日本語にしぼり，学習者の作成した誤りのうち仮名表記誤りの検出，訂正，助詞誤りの検出，活用形誤りの検出の方法を研究している．

本稿では上記の検出・訂正に際して，外国人が理解しやすい形態で結果を出力するための形態素解析について報告する．

2. 外国人向け日本語教育の特徴

日本語学校において外国人の受ける日本語教育と，一般的に日本人の受ける国語教育とでは，その内容は大きく異なっている．外国人向けの初級日本語教育は，会話によるコミュニケーション能力を重点に，日本語能力を獲得することを目的としている．日本語の能力をある程度獲得していることが前提の日本人向けの国語教育に対して，外国人向けの初級日本語教育は，ひらがなやカタカナ，漢字の読み書きのような初歩のことなどまったく知識がないことが前提である．そのため，外国人向けの初級日本語教育においては，会話を学習するのに適した文法を用いており，国語教育のものとは異なる．

例えば，動詞の活用であれば，日本人は五段活用・上一段活用・下一段活用・サ行変格活用・カ行変格活用という方法で学んでいる．これに対して，外国人はグループという概念で学んでいる．五段活用はグループ 1，上一段活用・下一段活用はグループ 2，サ行変格活用・カ行変格活用はグループ 3 とされている。動詞の活用形であれば，日本人が学ぶものには未然形・連用形・終止形・連体形・仮定形・命令形の 6 種類の活用形があるが，外国人は辞書形・マス形・ナイ形・テ形・タ形ほか，33 種類の活用形を用いて学習している[1][2]．外国人向けの日本語教育は，日本人が学ぶ国語教育とは異なっている．

本稿では，外国人向け活用形を構成する形態素を結合した見せかけの形態素を，擬形態素と呼称する．

† 東京電機大学大学院 未来科学研究科，
Graduate School of Science and Technology for Future Life,
Tokyo Denki University

3. 外国人向け形態素解析の構成方法

3.1 日本語形態素解析器の利用可能性

本研究では，日本語形態素解析器 MeCab[3]を利用している．MeCab は形態素解析辞書に IPA 辞書[4]を利用している．

外国人向け初級日本語教育において使われる活用形に対して，MeCab を用いて形態素解析した例を表 1 に示す．

表 1　外国人向けの活用形の形態素解析例

活用形	活用変化	解析結果
辞書形	書く	書く　動詞,自立,*,*,五段・カ行イ音便,基本形,書く,カク,カク
マス形	書き	書き　動詞,自立,*,*,五段・カ行イ音便,連用形,書く,カキ,カキ
ナイ形	書か	書か　動詞,自立,*,*,五段・カ行イ音便,未然形,書く,カカ,カカ
テ形	書いて	書い　動詞,自立,*,*,五段・カ行イ音便,連用タ接続,書く,カイ,カイ て　助詞,接続助詞,*,*,*,*,て,テ,テ
タ形	書いた	書い　動詞,自立,*,*,五段・カ行イ音便,連用タ接続,書く,カイ,カイ た　助動詞,*,*,*,特殊・タ,基本形,た,タ,タ

現状の形態素解析器の形態素単位での解析結果の出力，および日本人向けに記述された素性は初級日本語学習者には非常に理解しづらく，動詞に関していえば擬形態素単位で解析結果の出力が行われることが望ましい．

MeCab は形態素解析の際，辞書に定義された形態素の表層形，左文脈 ID，右文脈 ID，単語コスト，および別途定義された各文脈 ID どうしの連接コスト表を用いて解析を行っている．これらパラメータを調整することで，活用形単位で形態素解析を行うことができる．

3.2 日本語形態素解析辞書の改変方法

擬形態素を構成する形態素のうち，先頭の形態素の左文脈 ID と，末尾の形態素の右文脈 ID をもち，適切なコスト値を付与した，複数の形態素を結合した擬形態素を MeCab の辞書に登録することによって，形態素解析器そのものの改変を行うことなく，形態素解析の結果の出力を改変することが可能になる．

4. 形態素解析辞書の改変

4.1 擬形態素の辞書への登録方法

擬形態素の左文脈 ID，右文脈 ID，単語コスト値を決定するために，まずは擬形態素を構成する形態素の情報を得る必要がある．

外国人向け動詞活用形の場合，活用形を構成する形態素には規則性があるため，その規則に基づき形態素を辞書より取得する．表 2 にテ形の例を示す．

表 2　活用形を構成する形態素の規則の例

動詞の活用	テ形の規則
五段活用	連用タ接続 + 助詞・接続助詞・テ
上一段活用	連用形 + 助詞・接続助詞・テ
下一段活用	連用形 + 助詞・接続助詞・テ
カ行変格活用	連用形 + 助詞・接続助詞・テ
サ行変格活用	連用形 + 助詞・接続助詞・テ

表 2 のように擬形態素を構成する形態素を取得し，その各形態素の情報によって，擬形態素が作成できる．

擬形態素の文脈 ID は以下のように設定する．

左文脈 ID:　前方に位置する形態素の左文脈 ID

右文脈 ID:　後方に位置する形態素の右文脈 ID

擬形態素の単語コストは，単語コストを C，擬形態素を構成する形態素の数を n，i 番目の形態素の単語コストをa_i，i 番目と i+1 番目の形態素間の連接コストをb_jとすると，以下のように求められる．

$$C = \sum_{i=1}^{n} a_i + \sum_{j=1}^{n-1} b_j$$

※連接コストは辞書の連接コスト表から得られる．

例として，五段活用の動詞のテ形「書いて」の擬形態素を作成するならば，「書いて」は形態素「書い」と形態素「て」により構成されており，左文脈 ID は「書い」の左文脈 ID である 687，右文脈 ID は「て」の右文脈 ID である 307 を設定する．

単語コスト C は，「書いて」を構成する形態素の数は「書い」と「て」の 2 つであるから，

C = 「書い」の単語コスト
　＋「て」の単語コスト
　＋「書い」と「て」間の連接コスト
　= 7883 + 5170 + (−7392) = 5651

※連接コストは形態素ペアの頻出度を表し，多くの場合負の値をとる．

このようにして，「書いて」の擬形態素情報が作成できる．

4.2 外国人向け活用形のコスト値の調整

活用形を擬形態素として登録するとき，4.1 節の方法だけでは，以下のように期待通りに形態素解析が行われない場合がある．

「彼は漢字を書けた」の形態素解析
期待する結果:　書けた (タ形)
実際の結果:　書け　(マス形) + た

マス形は～マスに続く形の活用形であり，「書け」のように連用形語幹のみで構成される．タ形である「書けた」はマス形「書け」と同じく連用形語幹を含むため，解析の際に競合が起こる．

この競合の解消のため，優先させたい活用形擬形態素に対しコスト値の調整を行う．コスト値は低いほど優先されるため，優先したい活用形擬形態素の単語コスト値を，任意の定数値で下げることで調整できる．

以下は外国人向け活用形の分類と競合の関係である．

① 語幹のみで構成される活用形(4 種)
② 語幹+助詞で構成される活用形(3 種, ①と競合)
③ 語幹+助動詞で構成される活用形(11 種, ①と競合)
④ 語幹+助動詞+助詞で構成される活用形(6 種, ①･③と競合)
⑤ 語幹+助動詞+助動詞で構成される活用形(6 種, ①･③と競合)
⑥ 語幹+助動詞+助動詞+助詞で構成される活用形(2 種, ①･③･⑤と競合)
⑦ 語幹+助動詞+助動詞+助動詞で構成される活用形(1 種, ①･③･⑤と競合)

5. 考察

4.1 節において，本稿では擬形態素を構成する形態素の取得に，活用形を構成する形態素の規則性を利用したが，規則の利用なしに単に形態素解析を行うことによっても形態素の取得は可能である．しかし，擬形態素にしたい形態素のかたまりを形態素解析にかけると，多くは前後の文脈に欠ける場合に，期待する解析結果が得られないことがある．本稿のように大量の擬形態素を作成する場合には，すべての解析結果が正しいかどうかを検証するのは現実的でないため，擬形態素の作成に形態素解析を利用する場合には，たとえば MeCab の制約付き解析のような，補助的な手段が必要と考えられる．

6. おわりに

本稿では，既存の辞書の情報を流用して擬形態素を辞書登録することで，形態素解析器に外国人日本語学習者向けの解析を行わせることが可能であることを示した．

本稿にて作成した擬形態素に，適宜必要な素性情報を付与することによって，さらに利用性が高まるものと考えられる．

謝辞

形態素解析器 MeCab，IPA 辞書の開発に携わった方々，研究協力していただいた杉野勝也氏に感謝いたします．

参考文献

[1] 田中よね, 牧野昭子, 重川明美, 御子神慶子, 古賀千世子, 石井千尋, “みんなの日本語 初級 I ”, スリーエーネットワーク, 1998.
[2] 田中よね, 牧野昭子, 重川明美, 御子神慶子, 古賀千世子, 沢田幸子, 新矢麻紀子, “みんなの日本語 初級 II”, スリーエーネットワーク, 1998.
[3] 形態素解析器 MeCab, http://taku910.github.io/mecab/.
[4] IPA 辞書, https://osdn.jp/projects/ipadic/.