ハングルテキストへの自動音素情報付与と学習システムへの応用

望月　源, 高野 寛大
東京外国語大学 外国語学部
e-mail:{motizuki, takano.tomohiro.kob}@tufs.ac.jp

1. はじめに

最近の韓流ブームなどもあり，日本人の韓国文化への関心が高まり，韓国語を学習し始める日本人も増大している．日本語と韓国語は系統的に同じ語族に属すとは認められていないものの似ている部分が多い．例えば, 両言語とも基本的には「S+O+V」の語順であり，冠詞や単数，複数の概念がなく，男性・女性・中性といった「性」の区別も存在しない．また，どちらにも助詞が存在する．

こうした類似点があるにも関わらず, 日本語話者にとって韓国語はあまり理解しやすいようには感じられない．その主な要因は，韓国語で使用される「ハングル」にあると思われる．

ハングルは, 他に類似のない韓国語独特の表音文字であり，ハングルに馴染みが薄く読み方を知らない多くの日本語話者にとっては，ラテン系アルファベットのように想像して発音するということができない．ハングルを読むためには，各文字の基本的な発音と，隣接する文字の組み合わせによって音が変化する多数の「発音の変化」規則を覚える必要がある．そのため，仮に韓国語に興味があったとしても,「ハングルで書かれた文をとりあえず声に出して読む」ということがすぐにはできない．一般に，文字を見てそれを音で表わせないもどかしさはストレスが大きく，必要以上の難しさを感じることにもなる．可能ならば，発音を知った上で，詳細を学習する方が，多くの学習者にとって学習意欲の持続や学習の楽しさの面からも好ましい．

本研究では，こうした要求を満すため,「発音の変化」に対応したハングルへの音素情報付与のための規則を作成する．また，発音記号に不慣れな日本語話者の便宜も考慮し，音素情報を日本語のカナに置き変える規則も作成する．作成した規則の精度を実験により評価する．さらにハングルを学習するシステムとして，作成したルールを用いて任意のハングルテキストに音素情報を付与するシステムを構築する．

2. ハングルについて

本節では, 簡単に韓国語とハングルについて説明する.

2.1. 文字の構造と発音

ハングルの 1 文字は 1 つか 2 つの子音字母と 1 つの母音字母の組合せからなり, 図 1 に示す 4 パターンのどれかの形をしている．各文字の 1 つ目の子音は「初声」，母音は「中声」，2 つ目の子音は「終声」と呼ばれる．この各字母の種類は，初声が 19 通り，中声が 21 通り，終声が 27 通りある[1]ので，図 1 の 1，2 のように終声のない文字も合わせると，ハングルとしての文字数は 19×21×(27+1)で 11172 になる．

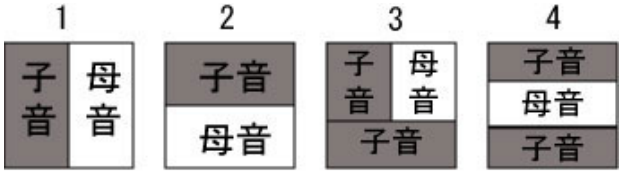

図 1 ハングルの構造

また，発音の方法によって，初声字母は「鼻音」「平音」「激音」「濃音」「流音」に分かれ，母音である中声字母は「単母音」「半母音+単母音」「二重母音」に分かれ，終声字母は,「口音」「鼻音」「流音」に分かれる[4][5]. ハングル 1 文字を発音する際には, 初声, 中声, 終声の順にそれぞれの字母が表す音を繋げて 1 音節で発音する．例えば,「감」は初声字母「ㄱ」，中声字母「ㅏ」，終声字母「ㅁ」からなりそれぞれの基本の音が平音の「k」，単母音の「a」，鼻音の「m」であるので,「kam(カム)」と発音する．

2.2. 発音の変化

ハングルでは，2 文字が連続する場合に，基本の音節から音が変化することがある．この「発音の変化」が起きるかどうかは，基本的に隣接する 2 文字の前の文字の「終声字母」と次の文字の「初声字母」の発音の仕方の組合せによって決まる．

例えば,「집념」(執念)の「집」は「chip」,「념」は「nyeom」であるが,「口音の終声字母の直後に，鼻音の初声字母が連接する場合，口音が鼻音に変化する口音の鼻音化」という規則があるため，口音の終声字母「ㅂ」と鼻音の初声字母「ㄴ」により「ㅂ」の「p」が対応する鼻音「m」に変化する．そのため全体として「chi**p**nyeom」から「chi**m**nyoem」に変化する．

[1]基本母音と合成母音，基本子音と濃音といった区別もある[3]が本研究ではすべてまとめて扱う.

終声と初声の間の発音の変化規則は，この他に「流音化」「激音化」「濃音化」「有声音化」があり，無音声の初声を挟んだ終声および中声の間の変化規則として「終声の初声化」「口蓋音化」がある．

2.3. 分かち書きと語彙

韓国語では文を記述する場合，分かち書きがされる．この分かち書きは日本語の文節にほぼ相当する．例えば，「私は」に相当する「저는（チョヌン）」や「私が」を意味する「제가（チェガ）」は，分かち書きされた1かたまりとなる．これを「単語」と呼ぶ場合もあるが，本研究では便宜的に「文節」と呼ぶことにする．

また，韓国語の語彙は，その成り立ちから，韓国固有の単語であり，日本語でいえば和語にあたる「固有語」，漢語にあたり，漢字で書き表すことができる「漢字語」，英語など他の言語を起源とする「外来語」の3種類が存在する．

3. 音素情報付与規則

本節では，ハングルテキストに対し計算機で自動的に音素情報を付与するために必要な規則と付与するための手法について説明する．また，音素表記をカナ表記にするための対応規則についても述べる．

3.1. 各字母と音素表記との対応規則

ハングルを構成する各字母には，対応する音素がある．そのため，初声，中声，終声の各字母1つずつに，基本の音と発音が変化した場合の音を表わす音素を対応させた次の形式のデータを作成する．

[ID ¥t 字母 ¥t P0 ¥t P1 ¥t P2 ¥t P3 ¥t P4 ¥t P5]

ここで，ID番号は各字母の通し番号で，初声はI00～I18(19通り)，中声はV00～V20(21通り)，終声はF00～F27(28 通り)である．P0～P5 は各字母について発音される可能性がある音素であり(最大6つ)，P0が基本の音を意味する．例えば，初声の「ㄱ」と「ㄲ」は

ID	字母	P0	P1	P2	P3	P4	P5
I00	ㄱ	k	g	kk	k'	-	-
I01	ㄲ	kk	-	-	-	-	-

となる．

3.2. 発音変化の規則

2.2 節で述べたように，発音の変化は，基本的に，連接する2文字の「終声字母」と「初声字母」の発音の仕方の組み合わせによって決定される．この発音の仕方は，字母の種類によって区別可能である．計算機によって自動的に発音変化を扱うため，以下の手順で，可能なすべての字母の組み合わせについて，変化の有無を調べ規則を作成する．

- 終声字母と初声字母の組み合わせをリストアップ(28×19=532 通り)し，各組み合わせにおける発音変化の有無を調べる．
- 発音変化が起こる場合，3.1 節の「各字母と音素表記との対応規則」に対応させて，各字母の基本の音素P0が，何番に変化するかを記述する．
- 発音変化がない場合，変化なしとする．

調査の結果，発音変化を伴なう組み合わせは532通り中，250通りあった．

さらに，ハングルには発音しない初声字母「ㅇ」(無声音)もあるため，その場合に起こりうる発音の変化2種類と，中声「ㅢ」に対応した例外1つの計3つの規則(以下に示す)を追加する．

- 終声「ㄷ」+初声「ㅇ」(無声音)+中声「ㅣ」の場合，終声音[t]を[ch]に変更する．
- 終声「ㅌ」+初声「ㅇ」(無声音)+中声「ㅣ」の場合，終声音[t]を[ch']に変更する．
- 後の文字の中声が「ㅢ」の場合，中声音[eui]を[i]に変更する．

最終的に535通りの組み合わせについて，以下の形式で「発音変化の規則」を作成する．

[終声 ID，初声 ID，中声 ID ¥t 変化の有無 ¥t 変化後終声音素 ¥t 変化後初声音素 ¥t 変化後中声音素]

この規則により，例えば「한국」(韓国) は1文字ごとには，「한」(han) と，「국」(kuk) と読むが，連続する場合，終声字母「ㄴ(F04)」と初声字母「ㄱ(I00)」の組み合わせである規則，

F04，I00，－ ¥t 有 ¥t P0 ¥t P1 ¥t －

が適用され，I00のP0(k)をP1(g)に変化させるため，「hanguk」(ハングク)となる．

3.3. ハングルへの音素情報の付与

「各字母と音素表記との対応規則」および「発音変化の規則」を用いて次の手順でハングルに音素情報の付与を行なう．なお，本研究では，以下の発音の変化が起きるかどうかを調べる範囲は，分かち書きされた「文節」内に限定し，文節をまたいだ文字の連続は考慮しないこととする．

1．最初の文字を，初声，中声，終声の各字母に分解
2．発音変化と関連しない字母に基本の音素を付与
3．隣接する次の1文字を取り出し，各字母に分解
4．字母の組み合わせにより，発音変化の規則を参照し，発音変化がある場合は，その音素を付与，変化のない場合は基本の音素を付与
5．次の文字がある場合は2へ，ない場合は最後の文字の終声字母に対応する音素を付与

なお，ハングルを初声，中声，終声の各字母に分解す

るために,本研究ではUnicodeを利用する[1].Unicodeにおいて，ハングルの字母は次の規則性を持つ.

- 終声字母(28種類)は1文字ごとに変化する.
- 中声字母(21種類)は28文字(各終声字母との組み合わせ)ごとに変化する.
- 初声字母(19種類)は588文字(中声字母21種類と終声字母28種類の組み合わせ)ごとに変化する.

この規則性を利用すると，任意のハングル1文字をその文字コードから分解して，構成する各字母が何番目の字母であるか計算することができる.

3.4. 音素表記とカナ表記の対応規則

本研究では，ハングルの読みを日本語のカナで表すための規則として，アルファベット表記の音素をカナに対応させる規則も作成する.

日本語のカナは，母音のみか子音＋母音の形が考えられる．一方ハングルでは，終声で終わる場合，最後が子音になる．この点を考慮してハングルとカナを対応させる以下の４種類の規則を作成する．なお，字母が異なっても音素が同じ場合もあるため，各規則数は字母の組み合わせの数とは異なる.

- (ハングルの)「中声」とカナの「母音」との対応.
  中声の音素に対応した21種類のルールを作成する．例:「a」と「ア」,「ae」と「エ」,「oe」と「エ」
- 「初声＋中声」とカナの「子音＋母音」との対応.
  全組み合わせとして504種類のルールを作成する.
  例:「d+ae」と「デ」,「cch+wae」と「ッチュェ」
- 「終声＋無声音の初声＋中声」とカナの「子音+母音」との対応.
  全組み合わせとして546種類のルールを作成する.
  例:「b+ya」と「ビャ」,「lt'+yeo」と「ルティョ」
- 「終声」と音の近いカナとの対応.
  8種類のルールを作成する.
  例:「ng」と「ン」,「p」と「プ」

これらの規則により,例えば,「한국」(韓国)は「h(初声)a(中声)n(終声)g(初声)u(中声)k(終声)」という音素列であるが,[h+a]で「ハ」,[n]で「ン」,[g+u]で「グ」,[k]で「ク」と対応し,「ハングク」となる.

4. 実験

3.3節のハングルへの音素情報付与の精度を確かめるため，次の2つの実験を行なう.

実験1 人為的なハングル文字列への自動音素付与
実験2 新聞記事への自動音素付与

4.1. 実験1

3.2節で述べたように，本研究では，ハングルの終声字母と初声字母の全ての組み合わせ525通り中の250通りと,例外規則3通りを合わせた253通りについて,「発音変化あり」として規則を定めている．実験1では，この253規則が正しく動作するかを網羅的に調査するため,各規則がそれぞれ最低2回試されるように,506種類のハングル文字列データを人為的に作成し,音素情報の付与を行なう.

作成した506文字列データの総文字数は1560(1データ当り平均3.08文字)で，発音変化の起こる可能性のある文字境界数は1054(平均2.08)であった.実験の結果，発音変化が発生した回数は，798回(平均1.58)あり，すべてに対して音素情報が正しく付与された.

ここで用いた人為的データは，すべての規則が適用されることを目的に作成されたものであり，実際の単語として存在しない文字列も含まれている．しかし,実験結果から，少なくとも作成した規則自体は正しく動作していることが確認された.

4.2. 実験2

実験2では，ハングルが実際に使用されているデータとして新聞記事を用いて,音素情報の付与を行なう.

今回のデータはインターネットの新聞サイト[2]から,「経済」「政治」「社会」「芸能」「スポーツ」の各ジャンルから各10記事ずつ，計50記事を収集した．各ジャンル別の平均および全体での結果を表1に示す.

表 1　実験2の結果

	経済	政治	社会	芸能	スポ	全体
文数	12.4	10.8	12.7	10.7	8.6	11.0
文字数	587.7	596.0	587.8	451.9	437.7	532.2
文節数	188.3	191.5	196.5	155.9	141.5	174.7
境界数	401.3	404.6	394.2	298.7	297.0	359.2
変化数	195.7	195.5	189.3	135.1	129.3	169.0
変化率	48.8%	48.3	48.0	45.2	43.5	47.1
正解数	193.2	195.5	189.3	133.7	127.8	167.0
正解率	98.7%	98.1	99.4	99.0	98.8	98.8
失敗数	2.5	3.8	1.2	1.4	1.5	2.1

表中で,「文節数」とは分かち書きされた 1 かたまりを示し,「境界数」とは文字と文字の境界のことで発音変化の起きる可能性のある場所を示す.また,「変化数」は境界の内で発音の変化が起きた数であり,「変化率」は,境界数に占める変化数の割合を示す.例えば,経済のジャンルでは1テキストあたり平均401.3境界に対して 195.7 回の発音変化が起こり，その割合が48.8%であることを意味する.

今回の新聞記事データで発音の変化の起きる割合は，43.5%～48.8%(全体平均で約47%，17958回中の8449回)であり，比較的頻繁に発音の変化が発生することがわかる．この内，正しく音素情報を付与できた割合(正解率)は98.1%～99.0%(平均98.8%，8449回中8345回)であった．この結果から今回の実験では良い精度で音素情報の付与が行なえているといえる．

音素情報の付与が正しく行なえなかった104例は，いずれも「漢字語」への音素情報付与の間違いであった．漢字語の中には，まったく同じハングル表記であっても，標準とは異なる発音の変化を伴うものが存在する．例えば，「발달」という文字列には，終声「ㄹ」と初声「ㄷ」の組み合わせにより発音が「lt」から「ld」に変化する規則により「paldal」という音素が付与される．しかし，この文字列は，「発達」という漢字語が対応するため，発音が変化せず，「palttal」としなければならない．こうした漢字語に対する発音の変化にも規則性があるが，正しく規則を適用するためには，任意のハングルが漢字語であるかを調べ，漢字語である場合，具体的にどの語であるかを特定する必要がある．この問題の解決には，意味解析などのより高度な処理が必要であり，本研究では今のところ対応していない．より正確な音素情報を付与するために，今後の課題として対応する必要がある．

5. ハングル学習システムへの応用

以上述べてきた音素情報付与プログラムを利用して，任意のハングルテキストに対して，音素情報を付与するハングル学習システムを作成した．システムのインタフェースを図２と図３に示す．

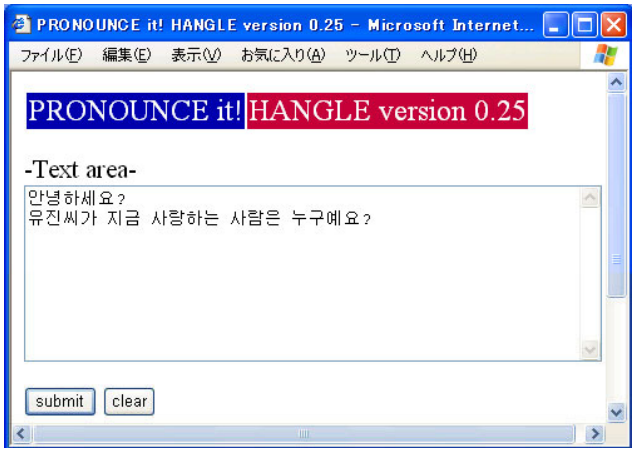

図 2　ハングルテキスト入力画面

本システムはWWWを利用し，利用者が図２の入力画面から，キーボードもしくは，他のファイルやWWWページからハングルテキストをはりつけ，「submit」ボタンを押すと，システムが自動的に音素情報を付与し，図３のような画面を表示する．

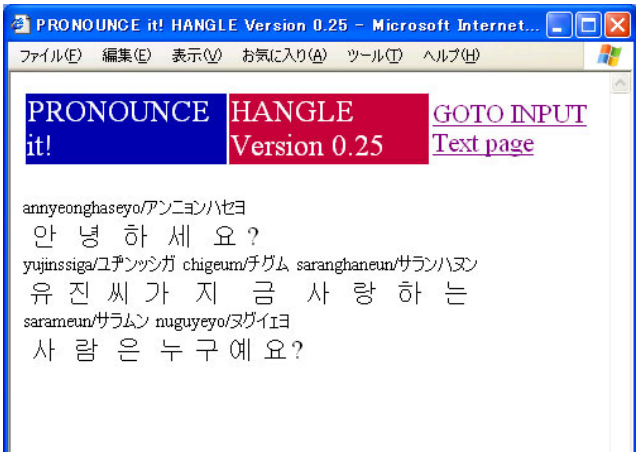

図 3　音素情報付与結果

音素情報を付与する際には，同時に3.4節で述べたカナ情報付与規則を利用し，日本語のカタカナによる情報も併せて表示している．このシステムにより，利用者は任意のハングルテキストを入力し，submitボタンを押すだけで，カタカナもしくは，音素情報によって読み方を知ることができる．

6. おわりに

本研究では，計算機によってハングルテキストに自動的に音素情報を付与するための規則を作成し，応用として，ハングルテキストに音素情報を付与するシステムを構築した．本手法により，比較的高い精度で音素情報の付与が行えた．今後の課題として，漢字語の発音変化に対応することと，今回扱っていない数字への対応を行い，より有用度をあげることがあげられる．

参考文献

[1] The Unicode Standard Version 4.0. **Hangle Syllables**. http:///www.unicode.org/charts/PDF/UAC00.pdf

[2] Digital Chosunilbo. **朝鮮日報**. http://www.chosun.com

[3] 金裕鴻, **韓国語がわかる。ハングルは楽しい!**, PHP研究所, 2004.

[4] 権在淑, **これからの朝鮮語**, 三修社, 1998

[5] 野間秀樹, 村田寛, 金珍娥, **ぷち韓国語**, 朝日出版社, 2004.